



DELHI UNIVERSITY  
LIBRARY

THE GIFT OF  
THE FORD FOUNDATION

**D.U.P. No. 93 - 0-8-15,000**

**RATAN TATA LIBRARY (DULS)**

**(Delhi School of Economics)**

**TEXT BOOK**

Cl. No. B280x

H8 51K959

Ac. No. 335283

This book should be returned on or before the date last stamped below. An overdue charge of 25 Paise per day will be charged for the first two days and 50 Paise for the third day the book is kept over.

[illegible]



*Practical*  
**BUSINESS  
STATISTICS**





# ***Practical*** **BUSINESS STATISTICS**

***FOURTH EDITION***

**FREDERICK E. CROXTON**

*Professor Emeritus of Statistics  
Columbia University*

**DUDLEY J. COWDEN**

*Professor of Economic Statistics  
University of North Carolina*

**BEN W. BOLCH**

*Assistant Professor of Economics  
Vanderbilt University*

© 1934, 1948, 1960, 1969 by

PRENTICE-HALL, INC., *Englewood Cliffs, New Jersey*

*All rights reserved. No part of this book may be  
reproduced in any form or by any means without  
permission in writing from the publishers.*

Printed in the United States of America

13-687798-2

Library of Congress Catalog Card No.:  
79-84459

Current printing (last digit):

10 9 8 7 6 5 4 3 2 1

PRENTICE-HALL INTERNATIONAL, INC.,  
*London*

PRENTICE-HALL OF AUSTRALIA, PTY. LTD.,  
*Sydney*

PRENTICE-HALL OF CANADA, LTD.,  
*Toronto*

PRENTICE-HALL OF INDIA PRIVATE LTD.,  
*New Delhi*

PRENTICE-HALL OF JAPAN, INC.,  
*Tokyo*

# Preface to The Fourth Edition

This text is designed for use by business administration and/or economics students engaged in a one- or two-semester course in basic statistics. The text has also been used by students whose primary interest lies in the other social sciences or in engineering.

The fourth edition differs in several respects from the third edition and from many other texts on business applications of statistical methods. Specifically:

1. Little or no attention is given to graphic or tabular presentation of data. It was felt that most current courses on business statistics have moved beyond the “chart drawing” exercises of past years.

2. Topics in elementary mathematics are covered in an appendix, and no attention is given to a presentation of elementary arithmetic.

3. There is an extended discussion of regression analysis, including the standard matrix algebra approach to the subject. It was felt that students who want to continue their course work in statistics need a “bridge” between the usual elementary presentation of regression and the presentation usually given in current works on econometrics. It is our hope that the undergraduate who has used this edition will find little difficulty in the transition to standard works on econometrics, such as those by Johnston, Goldberger, Christ *et al.*

4. There is an extended presentation of Bayesian inference in this edition together with some discussion of the similarities and differences between “classical” and “Bayesian” techniques.

5. Short problems are provided which are designed to test the student’s understanding of the content of the text and in some cases to extend the discussion.

6. In general much less attention is given to calculation in this edition than in previous ones. However, references on sources of computer programs are given which will carry out most of the calculations needed in elementary statistics. At Vanderbilt the mimeographed version of the text was used with success in a course which required that students do much of their calculation on a digital computer.

Many people have contributed to this edition of *Practical Business Statistics*. Ronald Wilder read all of the copy and made many helpful suggestions. Anne Bolch performed the clerical duties which make textbook writers glad that they are married. Many of our colleagues and students at Vanderbilt and the University of North Carolina offered criticism. In this respect we are especially indebted to John Pilgrim and the students of Statistics 251 at Vanderbilt. We are indebted to the literary executor of the late Sir Ronald A. Fisher, F.R.S., to Dr. Frank Yates, F.R.S., and to Oliver and Boyd, Ltd., Edinburgh, for permission to reprint all or portions of Tables III, IV, and V from their book *Statistical Tables for Biological, Agricultural and Medical Research*, 5th ed., 1957. We also wish to express gratitude for permission to reprint all or portions of Tables 8, 18, and 41 from E. S. Pearson and H. O. Hartley, eds., *Biometrika Tables for Statisticians*, Cambridge University Press, Vol. I, 1954.

Our thanks are also due Professor Larry E. Price, Division of Business, Georgia Southern College, for his helpful comments during the preparation of the manuscript.

This fourth edition of *Practical Business Statistics* was prepared by Ben W. Bolch.

*Nashville, Tennessee*

# Table of Contents

## 1

### **An Introduction to Statistics, 1**

- 1.1 Uses of Statistics in Business, 3*
- 1.2 Misuses of Statistics, 4*

## 2

### **The Frequency Distribution, 7**

- 2.1 Raw Data, 7*
- 2.2 The Array, 8*
- 2.3 The Frequency Distribution, 9*
- 2.4 Graphic Presentation, 11*
- 2.5 Determining the Number of Classes, 13*
- 2.6 Determining the Class Limits, 13*
- 2.7 Interpretation of Frequency Distributions, 16*
- 2.8 Percentage Frequency Distributions, 18*
- 2.9 Cumulative Frequency Distributions, 18*

### 3

#### **Averages, 22**

- 3.1 *Arithmetic Mean, 22*
- 3.2 *Two Mathematical Properties of the Arithmetic Mean, 24*
- 3.3 *Mid-range, 26*
- 3.4 *Median, 26*
- 3.5 *Mode, 27*
- 3.6 *Characteristics of the Mean, Median, and Mode, 28*
- 3.7 *Computations Using Grouped Data, 32*
- 3.8 *Other Means, 33*

### 4

#### **Dispersion, 39**

- 4.1 *Range, 40*
- 4.2 *Mean Deviation, 42*
- 4.3 *Standard Deviation  $SD$ , 43*
- 4.4 *Standard Deviation  $s$ , 45*
- 4.5 *Two Mathematical Properties of the Standard Deviation, 45*
- 4.6 *Efficient Computational Methods for  $s$ , 46*
- 4.7 *Relative Dispersion, 48*

### 5

#### **Shapes of Frequency Distributions, 51**

- 5.1 *Skewness and Kurtosis, 51*
- 5.2 *Standardized Distributions, 52*
- 5.3 *Moments, 54*
- 5.4 *Fisher's  $k$ -Statistics, 58*

## 6

**Probability and Some Discrete Probability Distributions, 60**

- 6.1 Definitions of Probability, 61*
- 6.2 Events and Sample Spaces, 63*
- 6.3 Probability Axioms, 64*
- 6.4 Probability and Independent Events, 65*
- 6.5 Probability and Dependent Events, 66*
- 6.6 Generalization and Extensions of the Formulas, 67*
- 6.7 Discrete Probability Distributions, 70*
- 6.8 The Binomial Distribution and its Parameters, 72*
- 6.9 The Poisson Distribution and its Parameters, 76*
- 6.10 Other Types of Discrete Probability Distributions, 78*
- 6.11 Bayes' Theorem, 78*

## 7

**The Normal Probability Distribution, 81**

- 7.1 Normal Curve as Limiting Form of Other Distributions, 82*
- 7.2 Probability and the Normal Curve, 89*

APPENDIX: Using the Normal Curve to Describe an Observed Frequency Distribution, 93

## 8

**Introduction to Statistical Inference, 100**

- 8.1 Point Estimators of the Population Mean and Variance, 102*
- 8.2 Point Estimators of the Population Proportion Defective, 106*



- 8.3 *Some Qualities of a Good Estimator, 106*
- 8.4 *Variance and Standard Error of the Mean, 108*
- 8.5 *Variance and Standard Error of the Sample Proportion, 109*
- 8.6 *Interval Estimates, 109*
- 8.7 *Tests of Hypotheses, 110*

## 9

### **Sampling Design, 113**

- 9.1 *Some Basic Ideas, 113*
- 9.2 *Methods of Sampling, 116*
- 9.3 *Sampling Designs, 119*

## 10

### **Tests of Hypotheses and Confidence Limits for the Arithmetic Mean:**

#### **Population Variance Known or Specified, 127**

- 10.1 *Two-sided Test, 127*
- 10.2 *One-sided Test, 130*
- 10.3 *The Power of a Test Concerning  $\mu$ , 132*
- 10.4 *Confidence Limits, 139*
- 10.5 *On Setting Alpha, 141*

#### **APPENDIX: Determination of Sample Size Through Risk Control, 142**

- A10.1 *One-sided Test Controlling both  $\alpha$  and  $\beta(\mu)$ , 143*
- A10.2 *Determination of Sample Size with Confidence Limits, 145*

## 11

### **Tests of Hypotheses and Confidence Limits for the Arithmetic Mean: Population Variance Unspecified, 147**

- 11.1 The t-distribution, 147*
- 11.2 Two-sided Test, 149*
- 11.3 One-sided Test, 152*
- 11.4 Confidence Limits, 153*
- 11.5 Hypotheses Concerning Differences between Two Populations, 154*

## 12

### **Tests of Hypotheses and Confidence Limits for Proportions and Standard Deviations, 160**

- 12.1 A Review and Extension of Some Previous Results Concerning Proportions, 160*
- 12.2 Hypotheses for Proportions, Infinite Populations, 162*
- 12.3 Hypotheses for Proportions, Finite Populations, 166*
- 12.4 Confidence Limits of a Proportion, 167*
- 12.5 Hypotheses Concerning Differences between Proportions, 170*
- 12.6 Tests of Hypotheses and Confidence Limits for the Standard Deviation, 171*
- 12.7 Hypothesis that Two Populations have the Same Standard Deviation, 172*

**APPENDIX: Discussion of the Variance of  $d$  and  $p$ , Finite Populations, 175**

## 13

### **Some Elements of Bayesian Decision Theory, 176**

*13.1 Two-sided Tests, 177*

*13.2 One-sided Tests, 184*

## 14

### **✓ Simple Linear Regression, 192**

*14.1 The Scatter Diagram, 192*

*14.2 The Least Squares Criterion, 195*

*14.3 The Normal Equations and their Solution, 196*

*14.4 Relationship of the Sample Regression Equation to the Population, 197*

*14.5 Tests of Hypotheses and Confidence Limits for the Slope and Intercept, 202*

*14.6 Estimation, 204*

*14.7 Alternative Statements of the Sample Regression Equation, 205*

*14.8 Regression Models, 206*

## 15

### **✓ The Correlation Coefficient, 208**

*15.1 The Standard Error of Estimate, 208*

*15.2 Two Alternative Concepts of the Correlation Coefficient, 209*

*15.3 Interpretation of the Correlation Coefficient, 213*

*15.4 Causation and the Correlation Coefficient, 215*

- 15.5 *Tests of significance, 216*
- 15.6 *Testing Other Hypotheses and Setting Confidence Limits, 218*
- 15.7 *Hypotheses Concerning Differences between Correlation Coefficients, 220*
- 15.8 *Correlation of Ranked Data, 222*
- APPENDIX: *Alternative Views of the Correlation Coefficient, 226*

## 16

### Multiple and Partial Correlation and Regression, 228

- 16.1 *A Three-variable Illustration, 228*
- 16.2 *The Normal Equations and their Solution, 231*
- 16.3 *Sources of Variation, 234*
- 16.4 *Multiple Correlation, 235*
- 16.5 *Partial Correlation and Relationships between Correlation Coefficients, 235*
- 16.6 *Alternative Approaches to Regression and Correlation, 240*
- 16.7 *Testing Significance, 241*
- 16.8 *Testing Other Hypotheses Concerning Partial Correlation, 245*
- 16.9 *Transformed Data, 245*
- 16.10 *The Multiple-partial Correlation Coefficient, 246*
- 16.11 *Adjusted Coefficient of Determination, 247*

### APPENDIX: Matrix Algebra Approach to Regression, 249

- A16.1 *Some Elements of Matrix Algebra, 249*
- A16.2 *Application of Matrix Algebra to Multiple Regression Analysis, 257*
- A16.3 *The Doolittle Computation Technique, 262*
- A16.4 *The Gauss-Markov Theorem, 263*

## 17

### **Tests of Homogeneity and Independence, 265**

- 17.1 Testing Homogeneity Using One-way ANOVA, 265*
- 17.2 Testing Homogeneity Using Two-way ANOVA, 272*
- 17.3 Testing Homogeneity Using Chi-square: Goodness-of-fit, 276*
- 17.4 Testing Homogeneity and Independence Using Chi-square: Contingency Tables, 279*

## 18

### **Index Numbers, 284**

- 18.1 Uses of Index Numbers, 285*
- 18.2 Problems in Index Number Construction, 286*
- 18.3 Index Number Symbols, 289*
- 18.4 Simple Index Numbers, 289*
- 18.5 Aggregative Price Index Numbers, 292*
- 18.6 Computation by the Method of Weighted Average-of-relatives, 298*
- 18.7 Changing the Base of an Index, 301*
- 18.8 Tests of Index Numbers, 303*

## 19

### **Time Series Analysis: The Secular Trend, 307**

- 19.1 The Problem of Time Series Analysis, 307*
- 19.2 Some Examples of Economic Time Series and Secular Trends, 309*
- 19.3 Test for Significance of Trend, 312*
- 19.4 Fitting a Linear Trend by the Method of Least Squares, 313*

- 19.5 *More Efficient Calculation of the Linear Trend by the Method of Least Squares, 315*
- 19.6 *Changing Units and Shifting Origin, 317*
- 19.7 *Higher Degree Polynomial Trends, 318*
- 19.8 *Growth Curves, 327*
- 19.9 *Moving Average as Trend, 338*
- 19.10 *Selection of Trend Type and Period to Which to Fit Trend, 342*

## 20

### **Time Series Analysis:**

#### **Seasonal, Cyclical, and Irregular Movements, 347**

- 20.1 *Calendar Variation, 349*
- 20.2 *A Conventional Method for Estimating a Stable Seasonal Component, 350*
- 20.3 *A Simple Method for Estimating a Moving Seasonal Component, 359*
- 20.4 *More Complicated Methods of Estimating a Moving Seasonal Component, 362*
- 20.5 *Estimating Cyclical Movements, 363*
- 20.6 *Irregular Movements, 367*

## 21

#### **Correlation of Time Series and Forecasting, 370**

- 21.1 *Correlation of Cyclical Relatives, 370*
- 21.2 *Forecasting a Series by Itself, 374*
- 21.3 *Forecasting a Series by Other Series, 379*
- 21.4 *Specific Historical Analogy, 380*
- 21.5 *Surveys of Plans and Opinions, 381*
- 21.6 *Cross-cut Economic Analysis, 381*
- 21.7 *Multiple Equation Models, 382*
- 21.8 *Judging the Accuracy of a Forecast, 384*

## Appendixes

1. *Values of  $Q(z)$  for Selected Values of  $z_Q$ , 391*
2. *Values of  $H(z)$  for Selected Values of  $z_H$ , 392*
3. *Values of  $z_Q$  for Selected Values of  $Q(z)$ , 393*
4. *Values of  $t_Q$  for Selected Values of  $Q(t \mid \nu)$ , 394*
5. *Values of  $F_Q$  for Selected Values of  $Q(F \mid \nu_1, \nu_2)$ , 395*
6. *Values of  $\chi_Q^2$  for Selected Values of  $Q(\chi^2 \mid \nu)$ , 398*
- 7-a. *Number of Combinations of  $N$  Things Taken  $n$  at a Time: Binomial Coefficients, 401*
- 7-b. *Values of  $e^{-a}$  for Selected Values of  $a$ , 402*
8. *Charts for Obtaining Confidence Limits for  $P$  in Binomial Sampling, Given  $d$  and  $n$ ; ( $p = d/n$ ), 403*
9. *Factors for Obtaining Unbiased Estimates of  $\sigma$ , and for Obtaining Control Limits for Ranges, when Sampling from a Normal Population, 405*
10. *Values of  $r_Q$  for Selected Values of  $Q(r \mid \nu)$  when  $\rho = 0$ , 406*
11. *Values of  $z_r$  for Selected Values of  $r$ , 407*
12. *Random Numbers, 408*
13. *Common Logarithms, 409*
14. *Squares, Square Roots, and Reciprocals, 414*
15. *Sums of Squares of Natural Numbers, 430*
16. *Flexible Calendar of Working Days, 431*
17. *A Review of Some Topics in Elementary Mathematics, 433*

## Index, 437

***Practical***  
**BUSINESS**  
**STATISTICS**





# I

## An Introduction to Statistics

As is the case in many disciplines, it is difficult to find a universally accepted definition of the science and subject matter of statistics. Many statisticians like to view statistics as a quantitative branch of scientific method which seeks to develop techniques for decision making under conditions of uncertainty. A slightly more concrete definition might state that statistics is a quantitative science that has as its goal the development and use of techniques for making valid inferences or judgments about populations on the basis of incomplete information. To the statistician the term *population* is simply a useful means of denoting the totality of the set of objects currently being considered. Hence a population might consist of all workers in a plant, all items produced by a machine in a given day, all items that could conceivably be produced by a given process, or all shares of stock traded on the New York Stock Exchange in a given month.

At first glance it would seem rather easy to make valid inferences about populations by the simple expedient of counting or measuring all of the members of that population with regard to the property in question. For example, it would be fairly easy to determine the “average” age to the nearest birthday of all employees of a firm that employed only 25 persons by simply asking each person to write down his age, adding together these ages, and dividing the resulting sum by 25. The figure obtained would be a type of average known as an *arithmetic mean*. On the other hand, it would be a good deal more difficult to find the average age of all persons living within the boundaries of a state such as New York. It is with regard to problems

of this latter type, where complete enumeration of the population is either impossible or impracticable, that the statistician and his techniques of measurement and observation become most useful. A trained statistician would attack this larger and more difficult problem by noting that it was unnecessary to ask every person in New York to divulge his age. Rather, he would take an adequate and representative sample of this population and then determine the specified average within certain limits of accuracy. Notice that in the first problem, where the entire population was studied, the statistician was simply *describing* a population characteristic. In the second problem the statistician was seeking to *infer* a population characteristic from the properties of a sample. The statistician might even go further and examine some hypotheses concerning the average age of the residents of the state of New York. Examples of such hypotheses might be: "The average age has not increased since the last time it was measured," or "The average age is not greater than 30 years," and so on.

The task of the statistician is made easier because of the fact that even though no two objects or individuals are ever exactly the same, there is generally some degree of uniformity among the members of a particular population. For example, even though no two men are of exactly the same height, it is well known that if one measures the height of many men, there will be a particular height that will be most often observed. It will also be seen that there are fewer men who are extremely tall or extremely short than there are men whose height is close to the most frequently observed value. The tendency toward uniformity as well as variability seems to be a law of nature.

Since the statistician recognizes that there is uniformity as well as diversity in nature, he views natural laws as being average relationships, not invariable mechanical ones. A physicist would say that whereas the behavior of an individual molecule is quite unpredictable, the behavior of innumerable molecules is often quite predictable within negligible limits of error. Of course, the accuracy of a statistician's prediction depends not only upon the variability of the population under consideration, but also upon the number of observations upon which the statistician's average is based. Averages of samples from the same population tend to become more stable if based on a large rather than on a small number of observations. For example, the batting average of nearly any baseball player varies considerably from game to game. From week to week there is less variability, and from month to month, still less. Of course, some players show a gradual improvement or deterioration during the season, and some players tend to show cycles, perhaps having occasional batting slumps. This is not an exception to the tendency of averages to become more stable as the sample size increases. The *apparent* exception results from the change over time in the characteristic being sampled.

Most people intuitively know a good deal about the nature of statistics. When the results of an examination are posted, for example, most students

are primarily interested in their individual grade. Then they usually want to know the class average and the "spread" of the grades about the class average.

## **I.1 USES OF STATISTICS IN BUSINESS**

Since the material in this book is not organized according to business uses, but according to the nature of statistical methodology, it is desirable to have a brief summary of business applications of statistical methods.

There are three major functions in any business organization in which the statistical method is useful. The first of these is in the planning of operations that may relate to special projects or to the recurring activities of a firm over a specified period of time, such as a year. The second relates to the establishment of standards that may pertain to volume of sales, quality standards for manufactured products, standard rates of output per day, or standard operating or financial ratios. The third function is that of control, which is attained by comparing achievement with plans or standards, and, when the former is too far out of line with the latter, taking appropriate action. Appropriate action is discovering what is wrong and then correcting the trouble or in some cases revising the standards. Planning, standard setting, and control are separate concepts, but they are interwoven in practice. Budgetary control includes planning and control, whereas quality control includes standard setting and control.

In marketing, statistical methods are used in the conducting of sales analysis, market analysis, and marketing analysis. Sales analysis is the study of sales records, properly classified; with the aid of sales analysis results of sales campaigns can be evaluated, areas discovered where special attention is needed, and plans formulated. Market analysis entails the study of market areas in order to estimate potential sales. Marketing analysis involves the study and comparison of different marketing methods.

In production, perhaps the most important use of statistics is in statistical quality control. Statistical quality control is used to establish standards of quality manufactured products, to control the manufacturing process so that the standard of quality is maintained, and to give assurance that individual lots sold or purchased are of acceptable quality. Other uses of statistics in production include the conducting and evaluation of tests for new products, time and motion analysis, and the planning of the replacement of physical equipment.

In personnel administration, statistical methods are used to devise and determine the reliability and validity of various achievement and aptitude tests that are used for hiring and promotion purposes. Statistical techniques are also used to determine wage adjustments brought about because of price level changes.

In finance, statistical methods are used in the computation and comparison of various financial ratios that are derived from accounting and various other data sources. Statistical techniques are also widely used for the forecasting of economic conditions that are known to affect the borrowing needs and borrowing costs of the firm.

Statistical methods are useful in accounting in the evaluation of accounts receivable and in the performance of various other accounting chores. For example, if a company has 20,000 customers with charge accounts, a representative sample might be taken of 1000 customers. A comparison of the book value of these accounts with amounts verified by correspondence with the sampled customers will often enable the auditor to make an inference concerning the bookkeeping accuracy of the other 19,000 accounts. In addition, similar sampling techniques are used in periodic inventory taking. Statistical techniques are also useful in calculating reserves for depreciation and in adjusting for the effects of price changes on depreciation and other reserves.

Finally, statistical methods often form the basis of certain business activities for organizations dealing with insurance, investment management, and gambling. The study of the application of statistical methods to insurance, called actuarial science, is a highly specialized field requiring years of study for its mastery. The development and application of statistical methods are also closely associated with the analysis of games of chance.

## 1.2 MISUSES OF STATISTICS

The famous quotation, "There are lies, damn lies, and statistics," points out very clearly that in many instances statistical methods are misused so as to impart erroneous impressions. In fact, it has been reported that individual firms, unions and agencies of various types attempt to hire statisticians to produce to order figures that seem to "prove" some specific point.<sup>(1)</sup> Happily, most statisticians are of high integrity, but the student of business statistics should be well aware of some of the ways in which statistical findings can be misused.

One of the most common activities in statistical work is the collection of data. When statistical conclusions are presented, the source and accuracy of the data upon which these conclusions are based should always be questioned. In almost any reputable statistical presentation the source of the data upon which the presentation is based will be explicitly given, and the absence of such citation should instill immediate doubts in the mind of the reader concerning the validity of the conclusions being presented. Also involved in data collection is the question of the definition of the variables under

---

<sup>(1)</sup> See William W. K. Freeman, "Training of Statisticians in Diplomacy to Maintain Their Integrity," *American Statistician*, December, 1963, pp. 16-20.

consideration. For example, when someone makes the statement that "prices are going up," does he mean consumer prices, wholesale prices, or some other price level indicator? Further, does he mean that his price level indicator is standing at a higher level this year than it stood one year ago, five years ago, or at some other past period of time, or perhaps that prices will be higher next year?

Unrepresentative and inadequate samples also often impart erroneous or misleading statistical conclusions. A sample may be unrepresentative if it is so designed that it fails to give a good cross section of the population under study, or if the method of sample collection is such that certain classes fail to respond in proper proportion. An inadequate sample is one that is so small in absolute size that very little reliability may be placed on the conclusions drawn from the sample. A study that caused considerable furor at the time it was published was Dr. Alfred C. Kinsey's *Sexual Behavior in the Human Female*. One criticism was that the sample used was only 5940 women, or less than one-hundredth of one percent of the female population. The percentage of the population that is included in the sample is of minor importance, provided that the sample is large enough in absolute size and designed in such a way that it provides a good cross section of the population. More important, it was argued by a religious leader that church members were inadequately represented. Because these and other presumably conservative classes were underrepresented, it was difficult to know how to interpret Dr. Kinsey's findings.

Unfair comparisons are also a source of trouble in statistics. There are so many ways of making unfair comparisons that it would be impossible to classify them all, but most unfair comparisons fall into four main categories.

1. *The making of absolute instead of relative comparisons.* It would be unfair to say that people are safer in an automobile than they are at home simply because more accidents happen in the home than in automobiles. What should be reported is the number of accidents that happen in both places relative to the number of man-hours spent in the home and in the automobile.

2. *The failure to classify data.* Politicians sometimes point to the large number of murders in the United States as verification of the need for better police protection of streets. What is not mentioned is the large number, perhaps the majority, of murders committed in the private home.

3. *The failure to allow for changes in composition.* It would be unfair to compare the death rates from heart disease for two states when the average age of the people in the two states is greatly different. Therefore, death rates are often standardized for age distribution.

4. *The use of a misleading base.* When one is making comparisons through time, the choice of the base year to be used in the comparison will usually have a great deal of effect on the resulting comparison. In the process of collective bargaining, a union representative might compare some index of

industry wage earnings in the current year with the same index in a past year when wage earnings were unusually high in order to show that wage earnings had increased very little or even decreased. On the other hand, the management representative might choose a base year when wage earnings were very low in order to show that wage earnings had increased a great deal. The proper base year, if it is possible to define one, would probably lie somewhere between these two extremes.

Forecasts made by lengthy extensions of the trend should be viewed with suspicion. Mark Twain, in a passage from *Life on the Mississippi*, illustrates the problem of unwarranted forecasts in the following way:

In the space of 176 years the Lower Mississippi has shortened itself 242 miles. This is an average of a trifle over one mile and a third per year. . . . Any person can see that 742 years from now the Lower Mississippi will be only a mile and three quarters long, and Cairo and New Orleans will have joined their streets together.

Another common statistical mistake is that of confusing association and causation. Simply because two phenomena are highly associated, it does not necessarily follow that one is the cause of the other. The disease malaria is so named because medical doctors once thought that it was caused by the bad air of swamps since the disease was invariably associated with swampy areas. The cause of this disease was not determined until it was experimentally shown that infection came from the bite of a female anopheles mosquito.

Computational errors and lies also often distort statistical findings. Even the greatest care cannot usually eliminate all errors in computation when the statistical study is of any but trivial magnitude. An analysis of the 1950 Census of Population will reveal that many 15-year old children were entering school for the first time and that the number of husbands living with their wives differed considerably from the number of wives living with their husbands. Presumably these and other interesting findings can be attributed to errors in the punching of computer cards. But even when computational errors are kept at a minimum by elaborate cross checks, results of samples may be distorted by overt or unintentional falsehoods. It is said that when many persons who voted Republican in the presidential election of 1960 were asked by political pollsters in 1964 how they voted in 1960, they reported that they had voted Democratic. Presumably, many people had the genuine conviction that they had voted Democratic because of remorse over the assassination of President Kennedy.

To summarize, a knowledge of the statistical method, its many applications, and the difficulties in the proper use of these applications are most important for the modern student of economics and business. In this time of rapidly expanding dissemination of quantitative information, the businessman who is unfamiliar with statistical methodology will find himself laboring under a severe handicap.

# 2

## The Frequency Distribution

In this chapter we will consider the technique of forming and plotting a frequency distribution, as well as some special charts that are useful for interpreting frequency distributions. We will also consider some ways in which frequency distributions may differ and what these differences may mean.

In this chapter and the three following, we confine our discussion almost exclusively to continuous quantitative variables. The following is a summary in outline form of types of variables.

1. *Qualitative variables.* These variables are the result of classification of attributes into mutually exclusive and exhaustive categories. Examples of qualitative variables are objects classified as defective or effective, persons classified as rich or poor, and persons classified as single, married, widowed, or divorced.

2. *Quantitative variables.* Quantitative variables may be discrete (discontinuous), continuous, or combinations of the two.

- a. Discrete variables take on only integral (whole number) values. For example: number of defects per unit; number of rich people in a sample of  $n$  persons.

- b. Continuous variables may be measured to an arbitrary degree of accuracy. For example, the weight of an object may be recorded in ounces, tenths of ounces, or thousandths of ounces. The result of the measurement is not necessarily a whole number.

### 2.1 RAW DATA

Statistical data may originally appear in a form such as that of Table 2.1. In this table are listed data representing the kilowatt-hours



**TABLE 2.1: KILOWATT-HOURS OF ELECTRICITY USED IN ONE MONTH BY 75 RESIDENTIAL CONSUMERS**

<i>Item number</i>	<i>Kilowatt-hours</i>	<i>Item number</i>	<i>Kilowatt-hours</i>	<i>Item number</i>	<i>Kilowatt-hours</i>
1	86	26	121	51	128
2	90	27	75	52	91
3	82	28	125	53	98
4	94	29	50	54	59
5	38	30	126	55	40
6	75	31	136	56	71
7	148	32	89	57	37
8	131	33	95	58	70
9	28	34	36	59	68
10	114	35	78	60	61
11	158	36	157	61	93
12	105	37	66	62	75
13	58	38	52	63	56
14	83	39	64	64	87
15	58	40	81	65	84
16	57	41	62	66	54
17	19	42	72	67	83
18	10	43	60	68	135
19	94	44	8	69	77
20	92	45	73	70	115
21	96	46	76	71	79
22	118	47	9	72	53
23	144	48	88	73	51
24	90	49	84	74	41
25	74	50	80	75	67

Source: Illustrative data.

of electricity used by each of 75 residential consumers of a power company. When data are in raw form as in this table, little information can be obtained by an inspection of them. To obtain such information as the greatest and smallest amount of electricity sold to a consumer, a careful examination involving every one of the 75 items is necessary.

## 2.2 THE ARRAY

As an aid in the simplification and interpretation of raw data a rearrangement of the raw data is often undertaken. When the raw data are arranged from the smallest to largest observation (see Table 2.2), or vice versa, the arrangement is often called an *array*. The array has certain distinct advantages over data in raw form. First, the range of consumption shown by the 75 readings is easily computed ( $158 - 8 = 150$ ), since the smallest and largest values can readily be seen; second, the tendency of the items to concentrate near some central value (such as 75 kilowatt-hours)

**TABLE 2.2: ARRAY OF KILOWATT-HOURS OF ELECTRICITY USED IN ONE MONTH BY 75 RESIDENTIAL CONSUMERS**

<i>Kilowatt-hours</i>	<i>Kilowatt-hours</i>	<i>Kilowatt-hours</i>
8	67	90
9	68	90
10	70	91
19	71	92
28	72	93
36	73	94
37	74	94
38	75	95
40	75	96
41	75	98
50	76	105
51	77	114
52	78	115
53	79	118
54	80	121
56	81	125
57	82	126
58	83	128
58	83	131
59	84	135
60	84	136
61	86	144
62	87	148
64	88	157
66	89	158

Source: Table 2.1.

may be observed; third, something can be seen of the distribution of the items between the upper and lower limits—notably that most of the items are between 36 and 136 kilowatt-hours, and very few are far below 36 or above 136 kilowatt-hours. In spite of these advantages, however, the array is cumbersome because there are still 75 separate items with which to deal. Obviously, the array would be even more cumbersome if we were dealing with several hundred items.

### 2.3 THE FREQUENCY DISTRIBUTION

Although some of the details shown by an array of individual items will be sacrificed, there is often much to be gained by summarizing the data into a *frequency distribution*. Such a frequency distribution is shown in Table 2.3.

In Table 2.3 the data are grouped into eight classes, each having an interval of 20 kilowatt-hours between the upper and lower limits of the class. Notice that each interval is not 19 hours, but 20 hours, because the actual class limits are as shown in column (2). Generally, when dealing with continuous

**TABLE 2.3: KILOWATT-HOURS OF ELECTRICITY USED IN ONE MONTH BY 75 RESIDENTIAL CONSUMERS**

(1)	(2)	(3)	(4)
<i>Stated class limits (consumption in kilowatt-hours)</i>	<i>Actual class limits (consumption in kilowatt-hours)</i>	<i>Mid-value of classes</i>	<i>Frequency (Number of consumers)</i>
5-24	4.5-24.5	14.5	4
25-44	24.5-44.5	34.5	6
45-64	44.5-64.5	54.5	14
65-84	64.5-84.5	74.5	22
85-104	84.5-104.5	94.5	14
105-124	104.5-124.5	114.5	5
125-144	124.5-144.5	134.5	7
145-164	144.5-164.5	154.5	3
Total	...	...	75

Source: Table 2.2.

variables, one takes measures of these variables to the nearest multiple of some acceptable unit. Recording the weight of an individual as 165 pounds generally means that the individual has an actual weight of something between 164.5 and 165.5 pounds. The measure of kilowatt-hours of electricity consumption in Table 2.1 is obtained after rounding to the nearest kilowatt-hour. Thus, the *actual* class limits, or class boundaries, given in Table 2.3 are not quite the same as the *stated* class limits, but are listed as being  $\pm 0.5$  units above and below the *stated* class limits. Notice that the stated class limits contain a break, or gap, of 1 kilowatt-hour between the stated upper limit of a given class and the stated lower limit of the following class, whereas the *actual* class limits do not contain such a break.

When one is using actual rather than stated class limits, the *class interval*, sometimes called the *width of class interval*, is easily found:

$$\text{Class interval} = \text{actual upper class limit} - \text{actual lower class limit}$$

For example, the interval of the first class may be found by noting that  $24.5 - 4.5 = 20$ .

The mid-value of any class may be found by averaging either the stated or actual class limits of the class. For the first class  $(5 + 24)/2 = (4.5 + 24.5)/2 = 14.5$ .

It will be noted that the class intervals of Table 2.3 are equal. Usually frequency distributions should be constructed with uniform class intervals. This procedure will facilitate the calculations to be described in the following chapters. Also, distributions with unequal class intervals are sometimes misinterpreted, especially when presented graphically.

## 2.4 GRAPHIC PRESENTATION

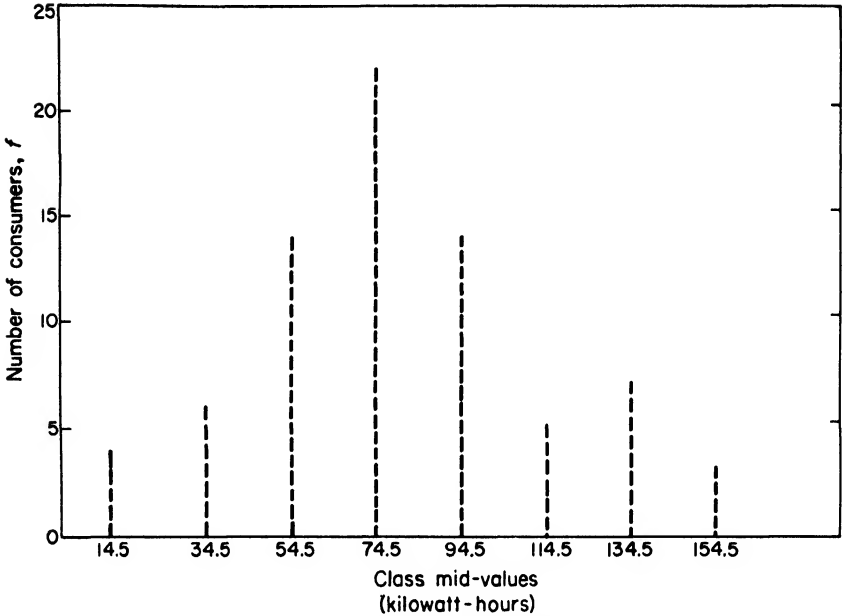
Chart 2.1 shows how the data of electricity consumption of Table 2.3 appears graphically. In the lower part of Chart 2.1 the vertical dimension of each bar represents the frequency, or number of occurrences in the class. The bars are, of course, all of the same width. This method of graphically portraying a frequency distribution is variously referred to as a *column diagram* or *histogram*.

There are several methods of constructing a histogram from a frequency distribution, and we will discuss one of these methods that is often convenient. First, as shown in the upper part of Chart 2.1, plot the mid-values of the frequency classes on the horizontal axis and plot directly above these points the number of occurrences in the class represented by the given mid-value. Next, mark off and label the upper and lower *actual* class limits for each class on the horizontal axis. Finally, construct the column for each class, using the actual class limits as bounds for the columns. The mid-values of the class intervals are generally not labeled on the horizontal axis of the finished histogram, since they can be found with ease from the presented actual class limits.

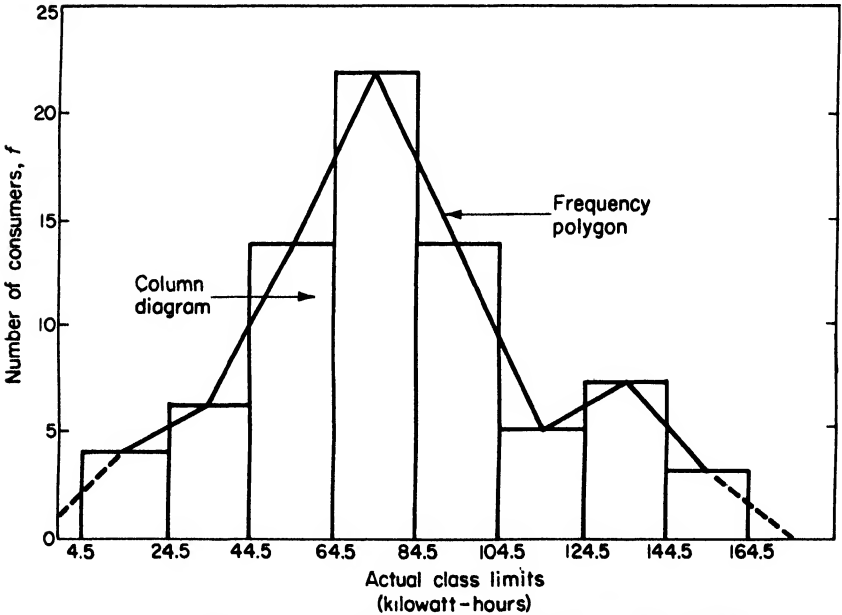
Instead of using a bar for each class, we may connect the mid-values of the class intervals with straight lines and form what is called a *frequency polygon*. The frequency polygon is often extended to the base of the diagram at distances of one-half of a class interval beyond the two end bars. To put it another way, the lines intersect the base at the mid-values of the next higher and lower class intervals. Thus in Chart 2.1 the polygon intersects the base at 174.5 and  $-5.5$  kilowatt-hours. There are several arguments in favor of such a procedure. First, there are no observations for these end classes, and therefore points should be plotted to indicate this fact. Second, the straight lines connecting the base points have no validity and are, therefore, plotted only for the convenience of the reader. Third, the chart is easier to understand if so plotted. Fourth, if the polygon is so plotted, the area under the polygon will be equal to the area under the histogram. However, the extension suggests the presence of information not actually found in the sample. Also, and perhaps more important, the extension to the left occasionally runs into negative values, as in Chart 2.1, when negative values are impossible.

Some statisticians use the column diagram for showing discrete data and the frequency polygon for showing continuous data. An objection to using the column diagram with discrete data is that there is the suggestion that class intervals are involved, which is often not the case with discrete data. For discrete data such as the distribution of the numbers of families having 1, 2,

**CHART 2.1: KILOWATT-HOURS OF ELECTRICITY USED IN ONE MONTH BY 75 RESIDENTIAL CONSUMERS SHOWN BY COLUMN DIAGRAM AND FREQUENCY POLYGON**



(a) Locating frequencies above class mid-values



(b) Final form of column diagram and frequency polygon

or 3 members, it is theoretically preferable to use vertical lines similar to those shown in the upper part of Chart 2.1. For popular presentation the column diagram is often used in preference to the frequency polygon; it is more striking and causes the values of the individual classes to stand out more clearly. For the purpose of comparing two sample frequency distributions, however, frequency polygons are more effective than are column diagrams.

Like many graphic representations of distributions generated in business and economics, Chart 2.1 is roughly bell-shaped, though it is not *symmetrical* about some central point. It is unusual to encounter frequency distributions that are exactly symmetrical, though nearly symmetrical distributions are often found.

## 2.5 DETERMINING THE NUMBER OF CLASSES

It is important that a frequency distribution be made with a suitable number of classes. If too few classes are used, the original data will be so compressed that little information will be available. If too many classes are used, there will be too few items in the classes, and the frequency polygon will be irregular in appearance. There are several rules of thumb available for determining the proper number of classes.<sup>(1)</sup> However, the number of classes is usually determined from problem to problem by a process of trial and error, i.e., by balancing information loss with irregularity of the frequency polygon until a pleasing compromise is reached in the eyes of the individual statistician. In practice about 12 or more classes are desirable for computational purposes, and the number of classes included is rarely less than four or five, or more than 20 or so.

## 2.6 DETERMINING THE CLASS LIMITS

There are several considerations to bear in mind when one is selecting the class limits.

<sup>(1)</sup> For example, a relationship has been suggested between the number of classes and the number of items to be classified when graphic presentation is desired.

<i>Number of items</i>	<i>Number of classes</i>	<i>Number of items</i>	<i>Number of classes</i>
5	2	100	10
10	4	200	12
25	6	500	15
50	8	1000	15

Source: Truman L. Kelley, "The Grouping of Data for Graphic Portrayal," Lecture at Cowles Commission Conference at Colorado Springs, Colorado, July 25, 1939.

It has also been suggested that the class interval not exceed one-fourth of the estimated population standard deviation. See George W. Snedecor, *Statistical Methods*, 4th ed. (Ames, Iowa: Collegiate Press, 1946), p. 170. Finally, according to the Sturges rule, the approximate number of classes,  $k$ , is given by:  $k = 1 + 3.3 \log n$ , where  $n$  is the number of observations and the logarithm is to the base 10.

1. *Have a representative mid-value.* As will be seen later, the mid-value of a class is used to represent all of the items in the class. Thus we must select the class limits so that the mid-values of the classes will coincide, so far as possible, with the concentrations of items that may be present. For example, in studying the meals sold by a cafeteria, it was found that a great many checks were multiples of 5 cents—that is, ended in 5 or 0. Consequently, the class intervals for a frequency distribution of the meal checks of this cafeteria read 8–12 cents, 13–17 cents, 18–22 cents, and so on, thus making the mid-values 10, 15, 20, and so on. In other words, the limits of the classes were chosen so that most of the luncheon checks would fall exactly on the mid-values of the classes.

Even though there are no points of concentration, class limits must be selected that are appropriate to the data. For example, if weights of metal castings are reported to the *nearest pound*, it is correct to write classes as shown in the table.

<i>Weight in pounds</i>	<i>Mid-value</i>
442–444	443
445–447	446
448–450	449
etc.	etc.

The class limits given above are, of course, stated and not actual class limits. The actual class limits would read 441.5–444.5, etc. If, however, the weights are given to the *last full pound*, the actual class limits would read as follows:

<i>Weight in pounds</i>	<i>Mid-value</i>
442 but less than 445	443.5
445 but less than 448	446.5
448 but less than 451	449.5
etc.	etc.

If several castings were weighed with the results shown in the first column below, the values would be recorded as shown in the two columns to the right.

<i>Actual weight</i>	<i>Rounded to the nearest pound</i>	<i>Rounded to the last pound</i>
443.501	444	443
443.674	444	443
444.499	444	444
444.730	445	444
449.427	449	449

Observe that the fourth item above would fall in the class "445-447" if recorded to the nearest pound,<sup>(2)</sup> but in the class "442 but less than 445" if rounded to the last pound.

2. *Avoid open-ended classes.* An open-ended class is one that includes all items smaller than some specified upper limit, or larger than some specified lower limit. Consider the distribution given in Table 2.4. If the first class in Table 2.4 had been labeled "under 50" and the last class had been labeled "80 and over" the two classes would be said to be open-ended. Since we would not know where the average grade in these classes was located, we would not be able to use these classes for computational purposes. Thus, we could not accurately estimate the average grade of the entire group of students.

There are times when open-ended classes are almost unavoidable because, without them, such a large number of classes would result that the frequency distribution would become unwieldy. Such a case might occur when there are some items with extremely large or extremely small values but when the majority of the other items are clustered in a relatively narrow range. For example, presentations of individual personal income in the United States usually have open-ended upper classes such as "\$10,000 or more."

3. *Class intervals should usually be uniform.* If not, the graphic presentation of the distribution might be misleading, because all class intervals are not the same (see Problem 6).

In cases where unequal class intervals are unavoidable because the distribution is markedly asymmetrical, it is best to report *adjusted frequencies* rather than frequencies. Adjusted frequencies are frequencies that have been adjusted for the class interval in which they lie. A standard interval is selected, and all class intervals are expressed as multiples of this interval  $c$ . The adjusted frequencies are then the original frequencies divided by  $c$ . Asymmetrical distributions can often be made symmetrical, or nearly so, by use of the logarithms of the class limits or of the square root of the class limits or by using class intervals that progress in some systematic manner.

TABLE 2.4: GRADES OF 50 STUDENTS ON A STATISTICS EXAMINATION

<i>Grade</i>	<i>Number of students (frequency)</i>	<i>Percentage of students (percentage frequency)</i>
40 but less than 50	5	10
50 but less than 60	10	20
60 but less than 70	20	40
70 but less than 80	10	20
80 but less than 90	5	10
Total	50	100

<sup>(2)</sup> Whenever a measurement falls on a class boundary or whenever it is impossible to determine on which side of a class boundary a measurement falls, a more accurate procedure is to split the frequency and place a half frequency in the class on each side.



## 2.7 INTERPRETATION OF FREQUENCY DISTRIBUTIONS

Frequency distributions may differ with respect to average value, dispersion, shape, or any combination of the three. Chart 2.2 compares two distributions that differ only with respect to average value. Distribution *A* is located farther to the right than distribution *B* and therefore has a larger average value. The location of the two arithmetic means is shown by a vertical line; the arithmetic means themselves are points located on the horizontal axis and are denoted by the symbols  $\bar{X}_A$  and  $\bar{X}_B$ . An average is often referred to as a measure of location.

**CHART 2.2: TWO DISTRIBUTIONS DIFFERING ONLY WITH RESPECT TO AVERAGE VALUE**

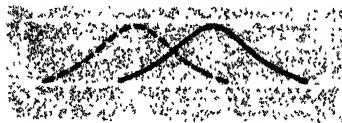
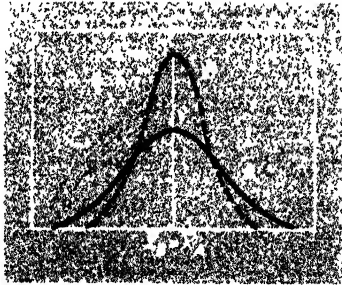


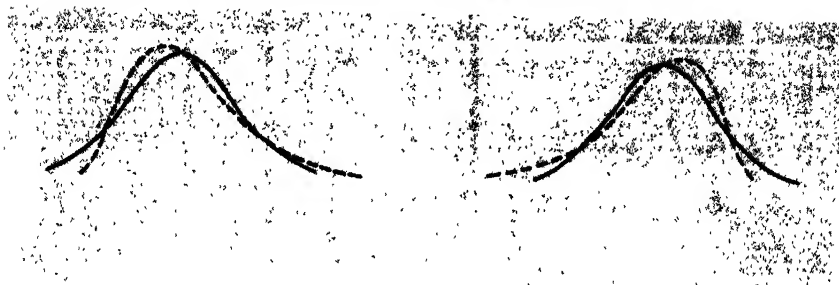
Chart 2.3 compares two distributions that differ only with respect to dispersion. Distribution *A* exhibits more dispersion; it is spread out more. Conversely, the values of distribution *B* are more uniform.

**CHART 2.3: TWO DISTRIBUTIONS WITH SAME AVERAGE VALUE BUT DIFFERENT DISPERSIONS**



The left side of Chart 2.4 compares two distributions that have the same degree of dispersion and the same average value but differ with respect to shape. Distribution *B* is bell-shaped and symmetrical about its arithmetic mean. In addition, distribution *B* is a normal distribution. The precise meaning of the term “normal” will be covered in a later chapter. On the other hand, distribution *A* has a “tail” pointing to the right. Distribution *A* is said to be *skewed* to the right (positively skewed). A positively skewed distribution has more extremely large values but fewer extremely small values than does a corresponding normal distribution. On the right side of Chart 2.4 distribution *B* is compared with distribution *C*, which is skewed to the left, or negatively. This condition indicates that distribution *C* possesses more

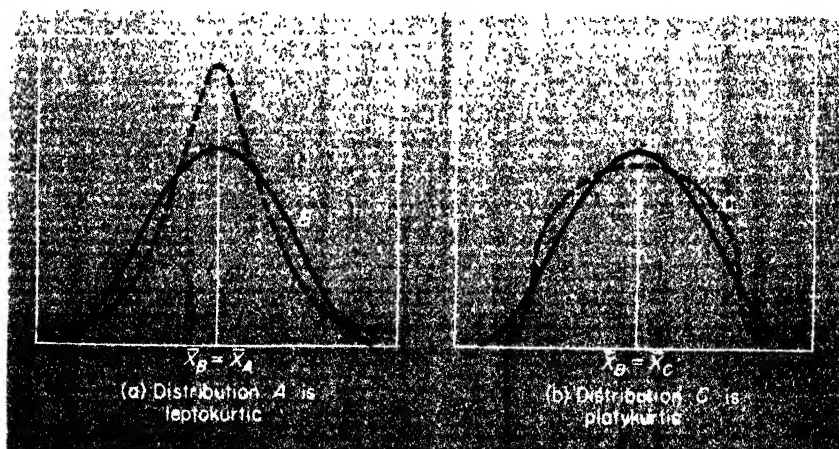
CHART 2.4: A POSITIVELY AND NEGATIVELY SKEWED DISTRIBUTION COMPARED WITH A SYMMETRICAL BELL-SHAPED DISTRIBUTION



extremely small values and fewer extremely large values than does distribution *B*, which is normal. Distributions that are skewed positively or negatively are said to exhibit *skewness*.

To illustrate another difference in shape, we see that the left side of Chart 2.5 compares two distributions with the same average value and the same amount of dispersion, but with shapes that differ in another respect. Curve *B* represents a normal distribution, whereas curve *A* has a smaller proportion of medium-sized deviations from the arithmetic mean, but a larger proportion of extremely large and extremely small deviations from the mean than does the normal distribution depicted by curve *B*. Such a distribution is said to be *leptokurtic*. Literally, leptokurtic means “with sharp curvature,” but this is a misnomer. Leptokurtic curves are usually, but not always, peaked. The right side of Chart 2.5 compares the normal curve *B* with a *platykurtic* curve, *C*. Distribution *C* has a smaller proportion than normal of deviations from the mean that are extremely small or extremely large, and a larger than normal proportion of medium-sized deviations from the arithmetic mean.

CHART 2.5: LEPTOKURTIC AND PLATYKURTIC FREQUENCY CURVES AND A NORMAL CURVE



Literally, platykurtic means "with wide curvature," but this also is a misnomer, since platykurtic curves are usually, but not always, flat-topped. Distributions which are leptokurtic or platykurtic are said to exhibit *kurtosis*. A normal distribution is said to be *mesokurtic*.

## 2.8 PERCENTAGE FREQUENCY DISTRIBUTIONS

The technique of transforming absolute frequencies into percentage frequencies is often useful when one wishes to compare two or more frequency distributions that are based upon different numbers of cases or different units of measurement, or both. The facilitation of comparison is accomplished because the percentage frequency distribution renders the area under the frequency polygons to be the same, i.e., 100 percent. The last column of Table 2.4 illustrates the formation of such a distribution. Notice that the total of this column, which represents the area under the percentage frequency polygon, is 100 percent.

## 2.9 CUMULATIVE FREQUENCY DISTRIBUTIONS

Understanding the content of a frequency distribution is often facilitated by cumulating the distribution. In Table 2.5 a frequency distribution relating to the length of life of light bulbs has been cumulated on a "less than" and also on an "or more" basis. The "less than" cumulative distribution shows the total number of light bulbs burning "less than" the number of hours indicated by the *upper* actual class limits of a given class. The "or more" cumulative frequency distribution shows the total number of light bulbs burning the number of hours indicated by the *lower* actual class limits of a given class "or more." For example, Table 2.5 shows that all, or 120, bulbs burned 250 hours or more and that all, or 120, bulbs burned less than 1650 hours.

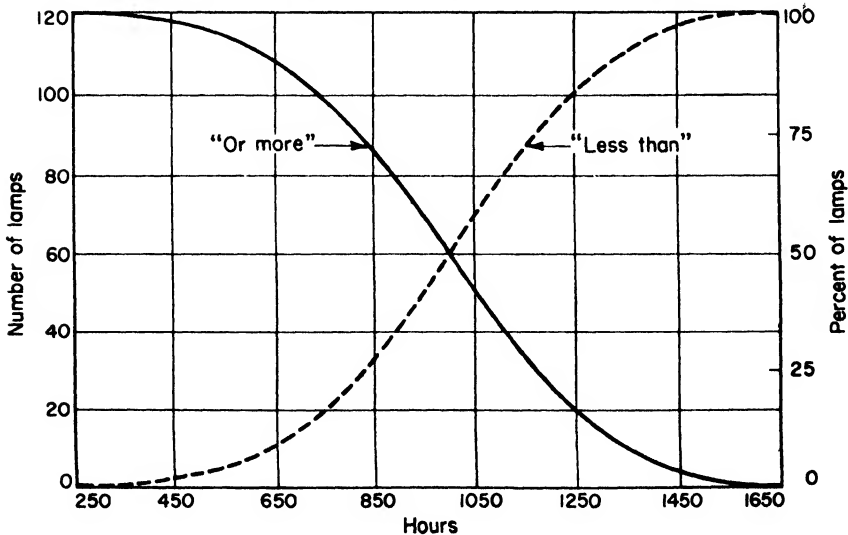
The two cumulative frequency curves, or *ogives*, are shown in Chart 2.6. The "less than" curve is shown by the dotted line and is found by plotting the "less than" cumulative frequencies above the upper actual class limits of the frequency classes. The "less than" curve intersects the base of the diagram at the lower actual class limit of the first class. The "or more" curve is drawn by plotting the "or more" cumulative frequencies against the lower actual class limits of the frequency classes. The "or more" ogive intersects the base of the diagram at the upper actual class limit of the last class.

It should be noticed that the "less than" and "or more" ogives display a characteristic "S" and "reverse S" shape, respectively. Also, the point at which these two curves intersect divides the frequency distribution into two equal parts. In Chart 2.6 the number of lamps is expressed in percentage

**TABLE 2.5: DISTRIBUTION OF LENGTH OF LIFE OF 120 60-WATT INCANDESCENT LAMPS AND CUMULATIVE FREQUENCY DISTRIBUTIONS**

FREQUENCY DISTRIBUTION		CALCULATION OF OGIVES		
<i>Life in hours</i>	<i>Number of lamps</i>	<i>Specified life in hours</i>	<i>Less than hours specified</i>	<i>Specified hours or more</i>
250 but less than 350	1	250	0	120
350 but less than 450	1	350	1	119
450 but less than 550	2	450	2	118
550 but less than 650	4	550	4	116
650 but less than 750	9	650	8	112
750 but less than 850	14	750	17	103
850 but less than 950	15	850	31	89
950 but less than 1050	26	950	46	74
1050 but less than 1150	19	1050	72	48
1150 but less than 1250	13	1150	91	29
1250 but less than 1350	9	1250	104	16
1350 but less than 1450	5	1350	113	7
1450 but less than 1550	1	1450	118	2
1550 but less than 1650	1	1550	119	1
		1650	120	0
Total	120	...	...	...

Source: Based on data from Electrical Testing Laboratories, New York City. Notice that each pair of entries in the two right columns sum to 120.

**CHART 2.6: 60-WATT INCANDESCENT LAMPS BURNING SPECIFIED HOURS OR MORE, AND BURNING LESS THAN SPECIFIED HOURS**

terms on the right side of the chart, and it should be noticed that the point of intersection of the two ogives is at 50 percent.

The construction of a "less than" frequency curve with a percentage vertical axis gives rise to the notion of a *percentile*, which is often used in test evaluation. For example, the "less than" ogive in Chart 2.6 shows that 25 percent of the light bulbs burned less than 843 hours. If the number 843 represented a raw test score, we would call it the twenty-fifth percentile, or the first quartile, meaning that 25 percent of the respondents made a test score this low or lower, while 75 percent made a test score this high or higher.

Quartiles and percentiles are often called *partition* values. They partition the frequency distribution into four parts and 100 parts, respectively. The intersection of the two ogives marks the value that divides the frequency distribution into two parts; this value represents the second quartile and fiftieth percentile and, as we shall see in the next chapter, the *median* of the distribution.

Partition values that divide a distribution into  $q$  equal parts are called *quantiles*. To compute a quantile one may proceed as follows:

1. Array the data from smallest to largest.
2. Compute  $k$ , the quantile item number:

$$k = \frac{p}{q} (n + 1)$$

where  $n$  is the number of items in the sample,  $p$  is the quantile number, and  $q$  is the number of parts into which the items are to be divided. The quantile item number  $k$  will often be a fraction.

3. If  $k$  is a fraction, compute a properly weighted average of the value represented by the integer part of  $k$  and the one following it. Thus, if  $k = 8\frac{3}{4}$  the desired quantile is  $(\frac{3}{4})X_8 + (\frac{1}{4})X_9$ . For most purposes (except when  $k$  ends in  $\frac{1}{2}$ ), it is satisfactory to round  $k$  to the nearest integer and consider the quantile to be  $X_k$  (see Problem 7).

---

## PROBLEMS

1. Distinguish between qualitative and quantitative variables. What is the difference between a quantitative variable that has been obtained by counting and one that has been obtained by measuring?
2. What are some practical criteria to be used in deciding what class limits should be used in constructing a frequency distribution?
3. Find the mid-values and actual class limits for the following distribution

of examination grades (fractional grades are possible). Plot the histogram and frequency polygon that results from your calculation.

<i>Classes</i>	<i>f</i>
40-49	2
50-59	4
60-69	7
70-79	12
80-89	3
90-99	2

4. Using the distribution given in Problem 3 above, calculate and plot the percentage frequency distribution.

5. Using the distribution given in Problem 3 above, calculate and plot "or more" and "less than" frequency curves.

6. Using the frequency distribution given in Problem 3 above:

*a. Group the first two and the last three classes; i.e., the stated classes now read: 40-59, 60-69, 70-99.*

*b. Plot the frequency polygon that results from this grouping and compare it with the frequency polygon that was calculated in Problem 3.*

*c. Adjust your grouped distribution for unequal class intervals; i.e., convert your frequencies to adjusted frequencies.*

*d. Plot the frequency polygon that results from step c.*

*e. Does the frequency polygon plotted in step b or the frequency polygon plotted in step d more closely resemble the frequency polygon plotted in Problem 3? Why?*

7. Establish the three quartile item numbers for the following data: 8, 15, 10, 10, 6, 7, 3, 4, 13, 14, 11. Determine the quartiles.

# 3

## Averages

An average is a single value within the range of the data that is used to represent all of the values in the series. Since an average is somewhere within the range of the data, it is sometimes called a measure of central value. Because the different values tend (in some sense) to cluster around this central value, an average is sometimes called a measure of central tendency. Thus in Chart 2.1 we saw a tendency for the items to concentrate most thickly around some value between 65 and 84 kilowatt-hours. Because an average tells us where a frequency polygon is located on the horizontal scale, an average is sometimes called a measure of location. Thus, it is apparent from inspection of Chart 2.2 that the solid curve is located farther to the right than the dotted curve, and therefore the average of the solid curve is greater than the average of the dotted curve.

In this chapter, we shall consider the following averages: arithmetic mean, mid-range, median, mode, geometric mean, harmonic mean, and quadratic mean. Relatively brief attention will be given to the last three mentioned.

We shall consider the concepts of these averages and illustrate how to compute them. We shall also introduce the use of the summation operator  $\Sigma$ . Students not familiar with the use of this operator are referred to Appendix 17 at the back of the text.

### 3.1 ARITHMETIC MEAN

Because it is used so frequently, the arithmetic mean is the most familiar average. It is the "average" of common parlance. The

arithmetic mean of a sample, generally indicated by the symbol  $\bar{X}$ , may be defined as

$$\bar{X} = \frac{\sum X}{n} \quad (3-1)$$

where  $n$  is the number of items in the sample.

Suppose that in a small factory drill press operators are receiving \$1.78, \$1.80, \$1.83, \$1.89, and \$1.95 per hour. The arithmetic mean would be

$$\bar{X} = \frac{\$1.78 + \$1.80 + \$1.83 + \$1.89 + \$1.95}{5} = \$1.85 \text{ per hour}$$

Sometimes a distinction is made between a "weighted" and an "unweighted" mean. The adjective unweighted is to be avoided, however, as all arithmetic means are weighted in some manner. A "simple" mean is one in which all the weights are the same. To obtain a weighted mean, one multiplies each observation by its weight, sums these products, and divides this quantity by the sum of the weights. Symbolically

$$\bar{X} = \frac{\sum WX}{\sum W} = \frac{W_1X_1 + W_2X_2 + \cdots + W_nX_n}{W_1 + W_2 + \cdots + W_n} \quad (3-2)$$

An instructor often assigns different weights to different parts of a course according to his opinion of their importance; for example: mid-term quiz, 1 point; laboratory, 1 point; final examination, 2 points. If the grades of a particular student are as follows: mid-term quiz, 67; laboratory, 89; final; examination, 81, the average grade is

$$\bar{X} = \frac{1(67) + 1(89) + 2(81)}{1 + 1 + 2} = \frac{67 + 89 + 162}{4} = 79.5$$

If the weights are relative weights, so that their sum is 1, computation of the arithmetic mean is simplified.

$$\bar{X} = \sum W'X \quad (3-3)$$

where  $\sum W' = 1$ .

Thus, we could write the weights: mid-term quiz, 0.25; laboratory, 0.25; final examination, 0.50. The computations are as follows:

<i>Part of course</i>	<i>Relative weight (<math>W'</math>)</i>	<i>Grade (<math>X</math>)</i>	<i><math>W'X</math></i>
Mid-term quiz	0.25	67	16.75
Laboratory	0.25	89	22.25
Final examination	0.50	81	40.50
Total	$\sum W' = 1.00$	...	$\bar{X} = \sum W'X = 79.50$



### 3.2 TWO MATHEMATICAL PROPERTIES OF THE ARITHMETIC MEAN

1. The sum of the deviations from the arithmetic mean is zero. Consider the five hourly wage rates used at the beginning of this chapter, and let  $x = X - \bar{X}$ . Then<sup>(1)</sup>  $\sum x = 0$ . Thus we have

$X$ (dollars)	$x = (X - \bar{X})$
1.78	-0.07
1.80	-0.05
1.83	-0.02
1.89	+0.04
1.95	+0.10
$\sum X = 9.25$ and $\bar{X} = 1.85$	$\sum x = 0$

2. The sum of the squares of the deviations from the arithmetic mean is a minimum. This statement means that when  $M$  is defined as  $\sum X/n$ , the quantity  $\sum (X - M)^2$  is at least as small as when  $M$  is defined in any other way.<sup>(2)</sup>

Consider our wage rate figures. Let us find the sum of the squares of the deviations from the mean.

$X$ (dollars)	$X - 1.85$	$(X - 1.85)^2$
1.78	-0.07	0.0049
1.80	-0.05	0.0025
1.83	-0.02	0.0004
1.89	+0.04	0.0016
1.95	+0.10	0.0100
Total	0	0.0194

<sup>(1)</sup> It is easy to prove that  $\sum x = 0$ . Since

$$\sum x = \sum (X - \bar{X}) = \sum X - \sum \bar{X}$$

and  $\bar{X}$  is a constant, we have

$$\sum x = \sum X - n\bar{X} = \sum X - n \frac{\sum X}{n} = 0$$

<sup>(2)</sup> For the student with a knowledge of elementary differential calculus, this theorem is easily proved. Let

$$f = \sum (X - M)^2$$

and set the first derivative with respect to  $X$  equal to zero to evaluate the minimum

$$\frac{df}{dX} = \sum (2X - 2M) = 0$$

Upon simplification we find that  $M = \sum X/n$ .

Next, consider the sum of the squares of the deviations from some other arbitrary point, say 1.83.

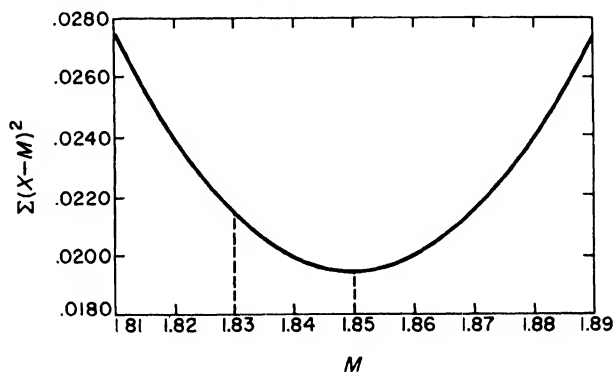
$X$ (dollars)	$X - 1.83$	$(X - 1.83)^2$
1.78	-0.05	0.0025
1.80	-0.03	0.0009
1.83	0.00	0.0000
1.89	+0.06	0.0036
1.95	+0.12	0.0144
Total	+0.10	0.0214

The results of these and other similar computations are plotted on Chart 3.1. It is apparent that when  $M = \bar{X} = \$1.85$ , the sum of the squared deviations is smallest. Therefore, we say that  $\sum x^2 = \sum (X - \bar{X})^2$  is a minimum.

As will be explained, the arithmetic mean is sometimes called a measure of central tendency in the sense that it may be thought of as the "center of gravity" of the data. In this context the use of the term "central tendency" is a misnomer, because it suggests a *degree* of concentration about some central value rather than the central value itself.

The arithmetic mean is always used when one is making an estimate of the mean of a normal population, for it is then the most reliable estimate of the population mean. This fact holds regardless of the sample size used to calculate the arithmetic mean. The precise meaning of the word "reliability" will be covered in a later chapter.

CHART 3.1: RELATIONSHIP BETWEEN  $M$  AND  $\sum (X - M)^2$



### 3.3 MID-RANGE

The mid-range, or center, is the arithmetic mean of the smallest item and the largest item. Thus if  $X_1$  is the smallest item and  $X_n$  is the largest,

$$MR = \frac{X_1 + X_n}{2} \quad (3-4)$$

For our five drill press operators

$$MR = \frac{\$1.78 + \$1.95}{2} = \frac{\$3.73}{2} = \$1.86$$

The mid-range is often used to compute average temperature or average price of a corporation stock, because maximum and minimum temperature or high and low price are of special interest.

It is hard to see how the mid-range can be thought of as a measure of central tendency, for there is no implication that the values in a sample tend toward the mid-range.

The mid-range is to be commended chiefly for its computational simplicity, though for extremely platykurtic distributions it is the most reliable estimate of the population mean. For normal or leptokurtic distributions it is extremely unreliable.

### 3.4 MEDIAN

The median is defined as a value that divides a series so that at least one-half of the items are as large as or larger than it is, and at least one-half the items are as small as or smaller than it is. For a series of values such as \$1.15, \$1.17, \$1.19, \$1.23, and \$1.29 it is clear that the median is \$1.19. If, however, we have an even number of items, for example, \$8.25, \$8.37, \$8.42, \$8.46, \$8.51, and \$8.52, our definition is satisfied by any value greater than \$8.42 but less than \$8.46. In such a case the convention is to take the median to be the mean of the two central values, in this instance, \$8.44.

The median is sometimes referred to as a *position average*. It must be obvious that the median cannot readily be determined for a series of figures unless they have been arrayed or organized into a frequency distribution. If an array is to be used, a little time may be saved by arraying merely the central part of a series. We must know how many smaller and how many larger items are present, but they need not be arrayed.

The median is said to be a measure of central tendency in the sense that an item larger than the median has the same probability as an item smaller than the median, provided there are no other items that have the same value as the median.

For samples from extremely leptokurtic distributions the median is the most reliable estimate of the population mean.<sup>(3)</sup> The median may be estimated graphically from a frequency distribution by use of cumulative percentage frequency curves. This technique was illustrated in Section 2.9. The median is located at the intersection of the ogives shown in Chart 2.6.

### 3.5 MODE

The mode is the value around which the items tend to concentrate. It is the most typical value and, therefore, the clearest example of a measure of central tendency. An item selected at random from some population has a greater likelihood of being the mode than any other single value.

As an example, consider the number of typing errors per day made on different days by a stenographer,

10, 10, 8, 9, 10, 10, 10, 11, 10, 9

It is almost imperative that data be arrayed in order to locate the mode, and preferably it should also be graphed. From the array

8, 9, 9, 10, 10, 10, 10, 10, 11

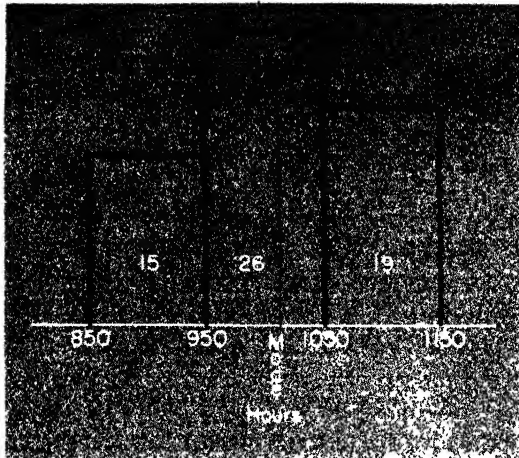
we see that 10 occurs most frequently and that other values occur with decreasing frequency the more they differ from 10. The mode is therefore 10.

The variable above is discrete, since there cannot be a fraction of an error. If it were continuous, we could say only that the mode is approximately 10, without further computation.

When one stops to consider that for a continuous variable there would not usually be two items of exactly the same size (if measurements are made with sufficient precision), it is apparent that our definition of the mode of a sample is somewhat vague. Because of this vagueness of concept, the best method of computing the mode of a sample is not obvious.<sup>(4)</sup> Before trying to estimate the mode, we ordinarily group the data. Consider Table 2.5, which shows the length of life of 60-watt incandescent lamps. Reference to that table reveals that the modal class, which is "950 but less than 1050 hours," has 26 observations. The next lower-valued class contains 15 observations, whereas the next higher-valued class has 19 observations. A diagram of these three classes only is shown as Chart 3.2. The mode is determined by the intersection of the two dotted lines. A perpendicular line from this point intersects the  $X$  axis at the mode. It appears to be about 1010 hours.

<sup>(3)</sup> Occasionally, when some of the extreme values are considered to be of doubtful validity, a specified number of the extreme values at each end is discarded, and the arithmetic mean of the remaining central values is computed. Such a measure may appropriately be called a *modified mean* if only a few items are excluded, or a *modified median* if only a few items are averaged.

<sup>(4)</sup> The mode of a probability distribution offers no conceptual difficulty. It is the value of abscissa which has the maximum ordinate.

**CHART 3.2: GRAPHIC ESTIMATION OF MODE.**

Source: Table 2.5.

In grouping the data one should make the class interval large enough so that the frequencies on each side of the mode get progressively smaller, at least for the two classes on each side of the modal class. Another problem is the location of the class limits. There are several alternative sets of classes for the frequency distributions formed from the same data and with the same class intervals. The particular way in which one groups the data may have an effect on the estimate of the mode.

Sometimes one encounters a distribution that has more than one mode or more than one frequency larger than those in their immediate neighborhood. Such data are usually heterogeneous; they are the result of mixing together two separate distributions. For such data, no average is meaningful. It is better to compute an average for each distribution.

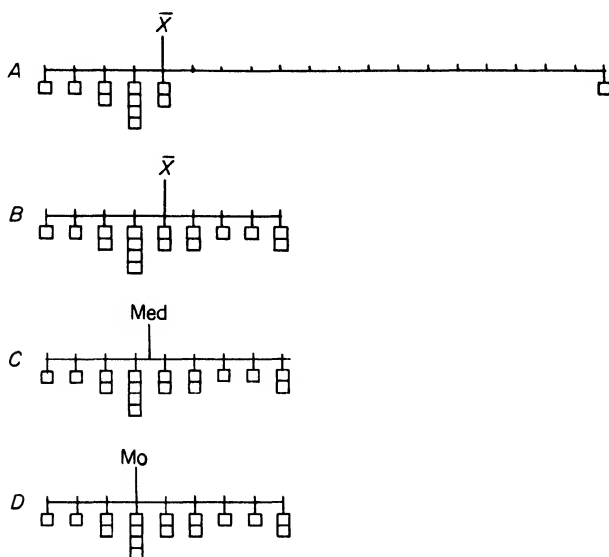
### 3.6 CHARACTERISTICS OF THE MEAN, MEDIAN, AND MODE

Sometimes the question is raised as to which of the three already described measures of central tendency is the "best." There is no simple answer to such a question. However, after considering the differing nature and behavior of the three measures, the student should know which measure or measures to use in a specific instance.

**Definitions.** It will be remembered that the three measures are based upon different concepts. The arithmetic mean is the sum of the values divided by the number of items. The median is the value that divides the series so that at least half of the items are equal to or greater than it, and at

**CHART 3.3: DIAGRAMMATIC REPRESENTATION OF ARITHMETIC MEAN, MEDIAN, AND MODE.**

(Distributions *A* and *B* are different, but have the same  $\bar{X}$ . Parts *B*, *C*, and *D*, show respectively  $\bar{X}$ , median, and mode for the same distribution.)



least half are equal to or smaller than it. The mode is the value around which the items tend to concentrate. (See Chart 3.3.)

**Mechanical Concepts.** The arithmetic mean is similar in concept to the idea of “center of gravity” as used in physics. Parts A and B of Chart 3.3 show two arrangements, each of which has the same mean. In part A the eight items to the left of the mean are offset by one item far to the right, so that the point of balance is at  $\bar{X}$ . In part B the same eight items to the left of the mean are offset by six items not so far removed as in part A. Parts B, C, and D show the same arrangement of items, but B shows the mean, C shows the median, and D shows the mode.

In terms of the frequency curve, the “center of gravity” idea of the arithmetic mean may not be easy for everyone to visualize. The median, however, divides the curve into two equal areas, and the mode is below the highest point of the smooth frequency curve. These averages are shown in Chart 3.4.

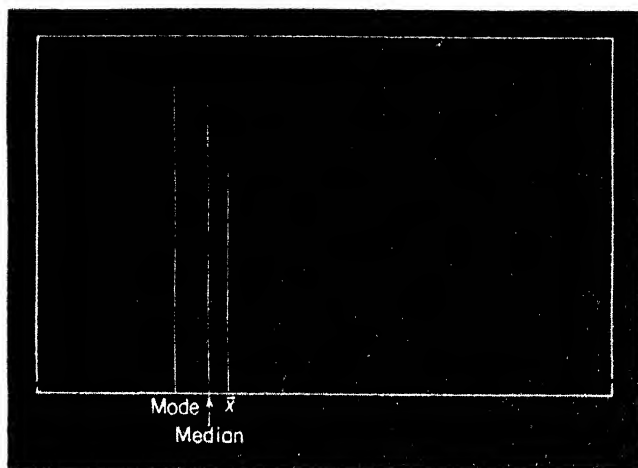
**Effect of Skewness and Extreme Values.** The frequency distribution of residential electrical consumption shown in Chart 2.1 is not symmetrical. It is slightly skewed to the right (positively skewed). The effect of this skewness is reflected in the values of the three measures of central tendency that we have discussed. It was found that

$$\bar{X} = 79.6 \text{ kilowatt-hours}$$

$$\text{Med} = 76.8 \text{ kilowatt-hours}$$

$$\text{Mo} = 74.5 \text{ kilowatt-hours}$$

**CHART 3.4: LOCATION OF ARITHMETIC MEAN, MEDIAN, AND MODE FOR A POSITIVELY SKEWED DISTRIBUTION.**



The arithmetic mean is the largest of the three values; the mode is the smallest. The reason for this difference lies in the fact that the arithmetic mean is influenced by the *value* of every item in the series, and the presence of a few large items increases the value of the arithmetic mean. These same items, however, have no more effect on the median than have the same number of items of *any* value greater than the median. The median is sensitive to the position of the values, but not to their size. The mode, as computed in this text,<sup>(5)</sup> is not at all affected by the extreme values in the series. When the skewness is to the left (negatively skewed), the mean is the smallest and the mode is the largest.

Because the arithmetic mean is sensitive to extremely large or extremely small values in a series, it may occasionally not be a typical value and may therefore be misleading. Suppose that six quarterly salaries are

$$\begin{array}{r}
 \$ 2000 \\
 2300 \\
 2400 \\
 2400 \\
 2500 \\
 10,000 \\
 \hline
 \Sigma X = \$21,600 \\
 n = 6 \\
 \bar{X} = \$ 3,600
 \end{array}$$

<sup>(5)</sup> Another method consists of determining the value below the highest point of a fitted curve. With such a procedure extreme values would have a slight effect on the value of the mode.

Because of the single extremely large salary the arithmetic mean is not typical and is therefore misleading.

**Effect of Kurtosis.** Occasionally a set of data may be found which has no mode. For example, a rectangular distribution, which is a type of platykurtic distribution, has no mode. A rectangular distribution is one in which the values are spaced at equal intervals, with each item occurring the same number of times. Consider the series: \$4, \$6, \$8, \$10, \$12. These data have no mode. The mean, median, and mid-range, however, are \$8. An extremely platykurtic distribution may be U-shaped. Such a distribution has a mean, median, and mid-range, but no mode.

**Algebraic Manipulation.** If we know any two of the three quantities  $\sum X$ ,  $n$ , and  $\bar{X}$ , the third quantity may be computed. If the arithmetic means of several series are to be averaged, this operation may be accomplished by using as weights the number of observations in each distribution (see Problem 6). The resulting mean of the means is also the mean of the combined distribution. Similar algebraic treatment for the median and mode is not possible.

**Mathematical Properties.** There are two important properties of the arithmetic mean. First, as was noted earlier,  $\sum x = 0$ . Second,  $\sum x^2 = \text{a minimum}$ , which means that the sum of the squared deviations of the items of a series about  $\bar{X}$  yields a smaller total than the sum of the squared deviations of the same items about any other value. The standard deviation, a measure of dispersion described in Chapter 4, is based upon  $\sum x^2$ .

The absolute sum (signs neglected) of the deviations of the items is a minimum about the median.

**Reliability.** The arithmetic mean is a more reliable estimate of the population mean (denoted symbolically as  $\mu$ ) than the median or the mode for a wide variety of populations. That is, there is less variability among arithmetic means computed from different random samples than among medians or modes. While this statement is not true for all types of populations, it is true for a normal population and for many others. This is an extremely important characteristic of the arithmetic mean; if the population is normal, the arithmetic mean is to be preferred to all other measures as an estimate of the population mean. For extremely leptokurtic distributions the median is the most reliable, whereas for extremely platykurtic distributions the mid-range is the most reliable.

**Necessity for Organizing Data.** It is not necessary to rearrange or group the raw figures to compute  $\bar{X}$ . In fact, it is not even necessary to



have the individual items. All that is necessary is to know the total  $\sum X$  and the number of items  $n$ . The median and mode cannot be easily computed unless the data have been arrayed or made into a frequency distribution.

**Effect of Open-end Classes.** Frequency distributions sometimes have open-end classes "Less than..." or "...or more." Because the mid-value of such a class is not apparent, the value of  $\bar{X}$  cannot be accurately determined for the series unless a note is appended to the table giving the total value of the items in the open-end class or classes. Open-end classes do not cause any difficulty in the location of the median and the mode, provided that the median or mode does not fall in the open-ended class.

### 3.7 COMPUTATIONS USING GROUPED DATA

**Arithmetic Mean.** In Table 2.1 data were shown of the consumption of electricity by 75 residential consumers. To obtain the arithmetic mean we may add the 75 values and divide by 75. The result is 79.7 kilowatt-hours.

When the data are in frequency distribution form as in Table 2.3, we are unable to sum the 75 individual items but must consider the distribution class by class. This we do by considering each class to be represented by its mid-value. These mid-values are then averaged, each being first multiplied by its respective frequency. The mid-values are our  $X_1, X_2, X_3, \dots$  values, and we use the following expression

$$\bar{X} = \frac{\sum fX}{n} \quad (3-5)$$

which is a special case of Eq. (3-2), where the weights are frequencies, and therefore  $\sum f = n$ . Table 3.1 shows the procedure for computing the arithmetic mean for the frequency distribution of sales of electric current. The value of  $\bar{X}$  is found to be 79.6 kilowatt-hours. The difference between this value and the more exact figure (79.7 kilowatt-hours) obtained from summing the 75 original items indicates that there was a slight loss of accuracy because of grouping the items into eight classes. The reason for this close agreement lies in the fact that, although each mid-value may be inaccurate as the representative value for a class, these inaccuracies *tend* to offset each other. Because of the central tendency within the distribution as a whole, the average of the values in a class below the class of greatest frequency usually will be greater than the mid-value of the class in question, whereas the average of the values in a class above the class of greatest frequency usually will be smaller than the mid-value of that class. Take, for instance, the 14 items in the class "45-64 kilowatt-hours." These 14 items may be identified in Table 2.2 and are found to average 56.8, which is *larger* than the mid-value 54.5. The 14

**TABLE 3.1: CALCULATION OF ARITHMETIC MEAN OF GROUPED DATA:  
KILOWATT-HOURS OF ELECTRICITY USED IN ONE MONTH BY 75  
RESIDENTIAL CONSUMERS**

<i>Consumption (kilowatt-hours)</i>	<i>Numbers of consumers f</i>	<i>Mid-value of class X</i>	<i>fX</i>
5-24	4	14.5	58.0
25-44	6	34.5	207.0
45-64	14	54.5	763.0
65-84	22	74.5	1639.0
85-104	14	94.5	1323.0
105-124	5	114.5	572.5
125-144	7	134.5	941.5
145-164	3	154.5	463.5
Total	75	...	5967.5

$$\bar{X} = \frac{\sum fX}{n} = \frac{5967.5}{75} = 79.6 \text{ kilowatt-hours}$$

Source: Table 2.3.

items in the class "85-104 kilowatt-hours" average 91.6, which is *smaller* than the mid-value 94.5. The tendency of the errors in the mid-value to offset each other will virtually never be perfect, but it will usually result in a value for the arithmetic mean of the frequency distribution that will closely agree with the arithmetic mean of the ungrouped data. The agreement will generally be closer for series showing continuous variation than for series showing gaps and concentrations, and it will usually be closer for symmetrical series than for skewed series. For very irregular or greatly skewed series the error attributable to grouping may be large.

**Median and Mode.** Graphic techniques for estimating the median and mode from a frequency distribution have been previously illustrated. Other methods, which utilize the numerical values of the frequencies and class limits directly, may be found in various texts.<sup>(6)</sup> For most practical purposes the graphic techniques previously illustrated are sufficiently accurate.

### 3.8 OTHER MEANS

A few other means are sometimes used: geometric, harmonic, quadratic. All of these means belong to the same family.

The geometric mean is the antilog of the arithmetic mean of the logs.

<sup>(6)</sup> See F. E. Croxton, D. J. Cowden, and S. Klein, *Applied General Statistics*, 3rd. ed. (Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1967), pp. 165-71.

The harmonic mean is the reciprocal of the arithmetic mean of the reciprocals.

The quadratic mean is the square root of the arithmetic mean of the squares.

**Geometric Mean.** The definition given above indicates that the geometric mean is the  $n$ th root of the product of the values. Thus, whereas the arithmetic mean of 4, 9, and 15 is

$$\bar{X} = \frac{4 + 9 + 15}{3} = 9.3$$

the geometric mean is

$$G = \sqrt[3]{4 \cdot 9 \cdot 15} = \sqrt[3]{540} = 8.1$$

Symbolically the geometric mean is

$$G = \sqrt[n]{X_1 \cdot X_2 \cdot \dots \cdot X_n} \quad (3-6)$$

For practical computation we write

$$\begin{aligned} \log G = \bar{X}_{\log} &= \frac{\log X_1 + \log X_2 + \dots + \log X_n}{n} \\ G &= \text{antilog} \frac{\log X_1 + \log X_2 + \dots + \log X_n}{n} \end{aligned} \quad (3-7)$$

and for a frequency distribution

$$G = \text{antilog} \frac{f_1 \log X_1 + f_2 \log X_2 + \dots + f_n \log X_n}{n}$$

where the  $X$  values are the mid-values of the classes.

Referring to the numerical illustration above, we find that  $G$  is smaller than  $\bar{X}$ . If one or more zeros are present,  $G = 0$ .

The essential difference between the arithmetic mean and the geometric mean may be made clear if we compare their mathematical characteristics.

1. While the sum of the deviations from  $\bar{X}$  is 0, the product of the ratios of the items to  $G$  is 1.
2. Series of the same number of items and having the same *total value* have the same arithmetic mean, whereas series of the same number of items and having the same *product* have the same geometric mean.

**Applications of the Geometric Mean.** Three applications will be mentioned.

1. Averaging rates. Consider the following data:

<i>Company</i>	<i>Net worth</i>	<i>Debt</i>	<i>Ratio of net worth to debt</i>	<i>Ratio of debt to net worth</i>
A	\$2500	\$1000	2.5	0.4
B	\$1000	\$2000	0.5	2.0

The arithmetic mean of the two ratios of net worth to debt is 1.5. But the arithmetic mean of the ratios of debt to net worth is 1.2, and  $1.5(1.2) > 1$ . This apparent absurdity arises because we failed to weight the ratios properly when we averaged them. Weighting the net-worth-to-debt ratios by using the denominators as weights gives

$$\frac{2.5(\$1000) + 0.5(\$2000)}{\$1000 + \$2000} = 1.167$$

which is, of course, the same as dividing total net worth by total debt.

$$\frac{\$3500}{\$3000} = 1.167$$

Similarly, we may average the debt-to-net-worth ratios and obtain

$$\frac{0.4(\$2500) + 2.0(\$1000)}{\$2500 + \$1000} = 0.8571$$

or, using totals, we have

$$\frac{\$3000}{\$3500} = 0.8571$$

These two figures are consistent with each other in that  $1.167(0.8571) = 1.00$ . However, these averages did not assign equal weights to the two ratios. If we wish to assign equal weight to each of the ratios being averaged and at the same time obtain consistent results, we may use the geometric mean. For the net-worth-to-debt ratios

$$G = \sqrt{2.5(0.5)} = \sqrt{1.25} = 1.118$$

and for the debt-to-net-worth ratios

$$G = \sqrt{0.4(2.0)} = \sqrt{0.8} = 0.8944$$

The reader should not infer from the preceding discussion that the geometric mean of the ratios is the correct measure to use and that the ratio of total values is incorrect, or vice versa. The measure to use depends upon the purpose. If for a number of firms it is desired to establish a typical net-worth-to-debt ratio, which will be independent of the amount of debt or of net worth of the different firms, the geometric mean may be used. If it is desired to ascertain what the net-worth-to-debt ratio of a number of firms would be after consolidation, then the proper figure is obtained by taking the ratio of total net worth to total debt.

Other cases in which the geometric mean is used for averaging ratios will be considered in connection with correlation and index numbers.

2. Skewed distributions. Occasionally a frequency distribution is encountered that is skewed to the right, but if logarithms of the  $X$  values are used, with the class interval of the logs constant, the curve becomes symmetrical. In such a situation the geometric mean may be appropriate, and the

antilog of the mean of the logarithms is also the geometric mean. One type of data that often falls into this category is ratios.

3. Averaging rates of change. For example: If a man invests \$400 in the stock market, and at the end of one year it has grown to \$500, he has had a 25 percent profit. If at the end of the next year his principal has grown to \$676, the rate of increase is 35.2 percent for the year. What is the average rate of increase of his principal during the two years? This rate may be obtained by the geometric mean. The average ratio is

$$\sqrt{1.25(1.352)} = 1.30$$

and the average rate of increase is therefore 30 percent.

The geometric mean may be used if more than two years are involved. Suppose we have the principal at five instants of time one year apart:  $X_0$ ,  $X_1$ ,  $X_2$ ,  $X_3$ ,  $X_4$ ; then

$$\begin{aligned} G &= \sqrt[4]{\frac{X_1}{X_0} \cdot \frac{X_2}{X_1} \cdot \frac{X_3}{X_2} \cdot \frac{X_4}{X_3}} \\ &= \sqrt[4]{\frac{X_4}{X_0}} \end{aligned}$$

From this formula we can see that we need to know only the principal at the beginning and the end of time under consideration. If the  $X$  values do not form a geometric progression, it is often better to fit an exponential curve to the data, as explained in a later chapter.

**Harmonic Mean.** The harmonic mean is defined as the reciprocal of the arithmetic mean of the reciprocals. Thus, for a simple harmonic mean

$$H = \frac{n}{\frac{1}{X_1} + \frac{1}{X_2} + \cdots + \frac{1}{X_n}} \quad (3-8)$$

For a "weighted" harmonic mean

$$H = \frac{\Sigma W}{w_1\left(\frac{1}{X_1}\right) + w_2\left(\frac{1}{X_2}\right) + \cdots + w_n\left(\frac{1}{X_n}\right)} \quad (3-9)$$

Although the harmonic mean is of limited usefulness, it is less affected by extremely large observations than any other average defined in this chapter. It is properly used to average rates where the weights are the numerators of the fractions used to compute the rates. The same result would be obtained by using the arithmetic mean and denominator weights (see Problem 4).

**Quadratic Mean.** It will be remembered that the geometric mean is the antilogarithm of the arithmetic mean of the logarithms, and the harmonic mean is the reciprocal of the arithmetic mean of the reciprocals.

Similarly, the quadratic mean is the square root of the arithmetic mean of the squares. Thus

$$Q = \sqrt{\frac{X_1^2 + X_2^2 + \cdots + X_n^2}{n}} \quad (3-10)$$

Use will be made of the quadratic mean in averaging deviations, rather than original values, when the standard deviation is computed in a later chapter.

**Averages of Means.** If several means are to be averaged, one should use the same method of averaging that was employed in computing the original averages. Thus one takes the arithmetic mean of several values of  $\bar{X}$ , the geometric mean of several values of  $G$ , the harmonic mean of several values of  $H$ , and the quadratic mean of several values of  $Q$ .

**Relative Magnitude of Different Means.** Provided all the  $X$  values are positive, and all of them are not the same, this relationship will always obtain:

$$Q > \bar{X} > G > H$$

## PROBLEMS

1. Give an example of
  - a. Five numbers that have no mode.
  - b. A position average.
  - c. A modified mean.
  - d. Five numbers where  $Mo > Med > \bar{X}$ .
  - e. Five numbers where  $Mo < Med < \bar{X}$ .
2. The number of persons living in a nation in year 1 was 2 billion, and in year 3 was 8 billion.
  - a. Estimate the number of persons in year 2 by using the arithmetic mean.
  - b. Estimate the number of persons in year 2 by using the geometric mean.
  - c. Why might the geometric mean be a more appropriate average than the arithmetic mean in this case?
3. Using the  $X$  items 1, 4, 2, 2,
  - a. Show that  $\Sigma x = 0$ .
  - b. Define some number  $M$  different from  $\bar{X}$  and show that  $\Sigma (X - M)^2 > \Sigma (X - \bar{X})^2$ .
  - c. Show that the product of the ratios of the  $X$  items to  $G$  is 1.
  - d. Form two series from the original one; i.e., 1, 4, and 2, 2. Show that the geometric mean of these two series is the same.
  - e. Show that the difference between  $G$  and  $\bar{X}$  is greater for the series 1, 4 than for the series 2, 2. Relate this to a mathematical property of  $G$ .

4. If you spend \$12.00 on \$3.00 shirts and \$10.00 on \$5.00 shirts, use the harmonic mean to find the average price per shirt.

5. Calculations using grouped data. Using the frequency distribution given in Problem 3, Chapter 2:

*a. Estimate the median and mode by graphic methods and calculate  $\bar{X}$ .*

*b. Under what conditions will the value of  $\bar{X}$  calculated from a frequency distribution be exactly the same as one calculated from the individual  $X$  values?*

6. If the mean of  $n_1$  numbers is  $\bar{X}_1$  and the mean of  $n_2$  numbers is  $\bar{X}_2$ , show that the mean of all numbers taken together,  $\bar{X}$ , is

$$\bar{X} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2}{n_1 + n_2}$$

Extend this result to the case of  $k$  different series of numbers.

# 4

## Dispersion

Chapter 1 pointed out that in nature chaos is as pervasive as order, and in statistical data, variability is as characteristic as similarity. It is only because of variability that we compute averages. We do not, for example, speak of the average number of days in a week. On the other hand, if there is too much variability among the data, an average is so unreliable that it is almost useless. When asked the average number of fish he caught during a day, the fisherman replied, "There ain't no average; it varies."

To characterize a sample merely by stating its average value is to give an inadequate description. Samples differ also as to their variability, or dispersion. Consider the data of Table 4.1, which shows the strength of the slotted end of valve caps made by the National Equipment Company and by a competitor. The mean strength of the National Equipment valve caps is 152.5 pounds; for the competitor the mean is 147.1 pounds. Therefore the National Equipment product is stronger. Looking more closely at the data, which are in arrays, we see that the tensile strength of the competitor's product varies from 65.7 pounds to 204.8 pounds. The National Equipment valve caps seem to be more uniform in quality, since its product only varies from 130.1 pounds to 180.7 pounds.

Usually a high degree of uniformity (small amount of dispersion) is a desirable quality. Mass production would usually be uneconomical if there were a large amount of variability in materials or manufactured parts, for interchangeability of parts is essential. On the other hand, although a uniformly mediocre athlete may never win a point in a track meet, an inconsistent performer of mediocre average ability will probably pick up several points during a season.



In this chapter we consider several measures of dispersion. But there are only two that are used much: the range and the standard deviation.

## 4.1 RANGE

The range is the difference between the largest value and the smallest value.

$$R = X_n - X_1 \quad (4-1)$$

For the illustration under consideration, the range of the National Equipment Company data is

$$R = 180.7 - 130.1 = 50.6 \text{ pounds}$$

whereas for the competitor's data it is 139.1 pounds. This difference indicates that National Equipment Company produces a more uniform product than its competitor.

Although it is the simplest measure of dispersion, the range has certain shortcomings. All of the shortcomings are attributable to the fact that the range does not utilize all of the information in the sample.

First, the range is less reliable than some other measures of dispersion, such as the sample standard deviation (to be discussed in Sec. 4.4). It varies too much from sample to sample taken from the same population. The range is unreliable because it is computed from the two most unreliable observations, i.e., the two most extreme observations, while being insensitive to the others in the distribution. Because the range is unduly influenced by unusual values, it would not ordinarily be used to describe a sample having one or a few unusual values.

Second, and less important, the range is extremely sensitive to the size of the sample. As the sample size is increased, the range tends to increase, though not proportionately. The range is said to be a *biased* estimate of the variability in the population as measured by the standard deviation of the population, which is defined in a later section. This means that if we took all possible samples of a given size from a population, calculated the range for each sample, and then averaged these ranges, the average range value would not be the same as the standard deviation of the population. The bias of the range can be corrected by multiplying it by a correction factor  $a_0$ , which varies with the sample size. That is,  $a_0R$  is an unbiased estimator of the population standard deviation. Values of  $a_0$  are given in Appendix 9.

In spite of its shortcomings, there are special situations where the range is satisfactory. When one is sampling from a normal population with a small sample size (say 12 or smaller), the quantity  $a_0R$ , which is an unbiased estimate of the population standard deviation, is nearly as reliable as the more laboriously computed standard deviation. When  $n = 2$ , they are equally

**TABLE 4.1: TENSILE STRENGTH OF VALVE CAPS MADE BY TWO COMPANIES (TENSILE STRENGTH IS IN POUNDS; DATA ADAPTED FROM CONFIDENTIAL SOURCE.)**

<i>Item number</i>	<i>Tensile strength <math>\bar{X}</math></i>	<i>Item number</i>	<i>Tensile strength <math>\bar{X}</math></i>
1	130.1	1	65.7
2	132.3	2	101.3
3	133.4	3	103.0
4	135.5	4	103.6
5	137.7	5	107.2
6	139.3	6	115.9
7	140.4	7	117.4
8	144.2	8	122.6
9	145.0	9	126.5
10	146.7	10	129.1
11	147.4	11	132.1
12	148.3	12	134.6
13	149.7	13	135.2
14	150.6	14	136.7
15	151.1	15	138.3
16	151.8	16	142.1
17	152.1	17	143.4
18	152.7	18	147.2
19	153.5	19	148.2
20	154.1	20	149.4
21	154.7	21	151.0
22	155.4	22	153.3
23	156.7	23	155.2
24	157.5	24	157.6
25	158.4	25	160.7
26	159.4	26	164.3
27	160.7	27	166.1
28	161.9	28	168.8
29	163.1	29	170.4
30	164.8	30	180.6
31	169.3	31	184.6
32	171.2	32	188.8
33	174.0	33	192.9
34	180.7	34	196.0
		35	200.4
		36	204.8
Total	5183.7		
Mean	152.5	Total	5295.0
		Mean	147.1

variable. In quality control, the range is customarily used for control charts, since the sample size is usually 4 or 5.

There are also certain types of data and certain purposes for which use of the range is appropriate. Among these are the range in temperature during the day or year and the range in stock prices during some period of time. In the latter case we learn the maximum profit that could have been made over the period by one purchase and one sale.

It is usually desirable to state the average for a set of data, as well as a measure of its dispersion. In the case of the range it is convenient and often desirable to state also the two values  $X_1$  and  $X_n$  from which the range, or the mid-range  $MR = (X_1 + X_n)/2$ , was computed.<sup>(1)</sup>

## 4.2 MEAN DEVIATION

The mean deviation, known also as the *average deviation*, is the mean of the absolute amounts by which the individual items deviate from the mean. It is computed by the following procedure.

1. Obtain the absolute deviations from the mean;<sup>(2)</sup> i.e.,  $|x| = |X - \bar{X}|$ . The vertical bars indicate that the sign of the deviations is disregarded; each of the  $n$  deviations is treated as if it were positive. See Table 4.2.

2. Sum the  $n$  deviations. Although  $\sum x = 0$ ,  $\sum |x| \neq 0$ . Table 4.2 shows that  $\sum |x| = 312.50$  pounds for the National Equipment data.

3. Divide the sum of the deviations by  $n$ .

Symbolically, these steps may be summarized as follows:

$$MD = \frac{\sum |x|}{n} \quad (4-2)$$

**TABLE 4.2: COMPUTATION OF  $\sum |x|$  FOR OBTAINING MEAN DEVIATION OF NATIONAL EQUIPMENT COMPANY DATA**

( $X$  = tensile strength in pounds;  $\bar{X} = 152.46$  pounds;  
data of Table 4.1)

Item number	$X$	$x = X - \bar{X}$	$ x  =  X - \bar{X} $
1	130.1	-22.36	22.36
2	132.3	-20.16	20.16
.	.	.	.
.	.	.	.
34	180.7	28.24	28.24
Total	5183.7	0	312.50

<sup>(1)</sup> Another easy-to-compute measure of dispersion that is sometimes used is the semi-interquartile range, or quartile deviation. It is most easily obtained by subtracting the lower quartile  $Q_1$  from the upper quartile  $Q_3$  and dividing by 2.

$$QD = \frac{(Q_3 - Q_1)}{2}$$

A characteristic of the quartile deviation is the fact that within  $\pm QD$  of the median  $Q_2$ , approximately 50 percent of the items are found. An advantage of this measure of dispersion is that it can be used with open-ended distributions.

<sup>(2)</sup> Occasionally the deviations are measured from the median, since  $\sum |X - M|$  is minimized when  $M$  is the median.

For the National Equipment Company data

$$MD = \frac{312.50}{34} = 9.2 \text{ pounds}$$

The mean deviation is a simple and easily understood measure of dispersion. Unlike  $R$ , it is affected by the value of each item. Although  $MD$  gives a reasonably good statement of the dispersion of a sample, it is not as frequently employed as is the standard deviation, which will be described in the next section.

There are several objections to using  $MD$ . First, it is awkward in that it is mathematically difficult to work with absolute deviations. Second, it is unreliable in that it varies too much from sample to sample taken from the same population. Third, it tends to increase with the size of the sample, though not proportionately and not so rapidly as the range. Finally, it is a biased estimator of the population standard deviation, which is the most often used measure of population dispersion.

### 4.3 STANDARD DEVIATION $SD$

The standard deviation of a sample  $SD$  is similar to the mean deviation in that it considers the deviation of each  $X$  value from  $\bar{X}$ . However, instead of using the absolute values of the deviations, it uses the squares of the deviations. These are summed, divided by  $n$ , and the square root extracted.

Sometimes it is convenient to think of the computations as involving three stages, each of which has a name, a symbol, and a formula.

<i>Name</i>	<i>Symbol</i>	<i>Formula</i>
Variation	$\Sigma x^2$	$\Sigma (X - \bar{X})^2$
Variance	$(SD)^2$	$\Sigma x^2/n$
Standard deviation	$SD$	$\sqrt{(SD)^2}$

$$SD = \sqrt{\frac{\Sigma x^2}{n}} \quad (4-3)$$

The computations for the National Equipment Company data are shown in Table 4.3. From these we obtain by stages the standard deviation.

<i>Name</i>	<i>Symbol</i>	<i>Computation</i>
Variation	$\Sigma x^2$	4770.34
Variance	$(SD)^2$	$4770.34/34 = 140.3$
Standard deviation	$SD$	$\sqrt{140.3} = 11.8 \text{ pounds}$

**TABLE 4.3: COMPUTATIONS OF  $\sum x^2$  FOR OBTAINING STANDARD DEVIATION OF NATIONAL EQUIPMENT COMPANY DATA**

( $X$  = tensile strength in pounds;  $\bar{X}$  = 152.46 pounds;  
data of Table 4.1)

Item number	$X$	$x = X - \bar{X}$	$x^2 = (X - \bar{X})^2$
1	130.1	-22.36	499.97
2	132.3	-20.16	406.43
.	.	.	.
.	.	.	.
34	180.7	28.24	797.50
Total	5183.7	0	4770.34

The concept of using sums of squares of deviations about the arithmetic mean of a distribution is very important. In later chapters extensive use will be made of this concept of dispersion. Because of its method of computation, the statistic  $SD$  is sometimes called the *root mean square deviation*. Most concisely of all, it may be defined as the quadratic mean of the  $x$  values.

The statistic  $SD$  is the most reliable of any measure of dispersion presented up to this time (except when  $n = 2$ , in which case  $SD = R/2$ ). It is usually true that a measure which utilizes all the information in a sample is more reliable than one which utilizes only part of it. Of the measures considered so far in this chapter, only  $MD$  and  $SD$  made use of each  $X$  value.

Although the statistic  $SD$  has the advantage of being in the same units of measurement as the original  $X$  values, the statistic  $(SD)^2$  is algebraically simpler to manipulate. Both  $SD$  and  $(SD)^2$  overcome the first two objections to the use of  $MD$ ; they are easier to manipulate mathematically, and they are reliable.<sup>(3)</sup>

Although reliable, both  $SD$  and  $(SD)^2$  are biased estimators of the population standard deviation and variance, respectively. Also, both  $SD$  and  $(SD)^2$  tend to get larger as  $n$  increases, though not proportionately. This tendency is not so marked, however, as is the case with  $R$ . Just as  $R$  can be corrected for bias by multiplying it by the correction factor  $a_0$ , so  $SD$  can be corrected for bias by multiplying it by the correction factor  $a_1$ . That is,  $a_1(SD)$  is an unbiased estimator of the population standard deviation. Values of  $a_1$  are given in Appendix 9. Also, as will be explained in the next section, the variance of a sample  $(SD)^2$  can be made into an unbiased estimator of the population variance by use of the correction factor  $n/(n - 1)$ . Thus,  $[n/(n - 1)](SD)^2$  is an unbiased estimator of the population variance.

<sup>(3)</sup> Also, for a normal population,  $SD$  is said to be a minimum variance estimator of the population standard deviation  $\sigma$ . Its variance is smaller than  $s$ .  $\bar{X}$  and  $(SD)^2$  are also joint maximum likelihood estimators of  $\mu$ , the population mean, and  $\sigma^2$ , the population variance, respectively.

#### 4.4 STANDARD DEVIATION $s$

In the last section we noted that although  $(SD)^2$  is considered less awkward and more reliable than  $MD$ , it is still a biased estimator of the population variance. However, the statistic

$$s^2 = \frac{\sum x^2}{n-1} \quad (4-4)$$

is an unbiased estimator of the population variance. The statistic  $s^2$  should usually be used to estimate the variance of a population. Thus,  $s^2$  eliminates all objections to  $MD$ . Also notice that

$$s^2 = \frac{n}{n-1} (SD)^2 = \frac{\sum x^2}{n-1}$$

which explains the correction factor given in the last section for  $(SD)^2$ . A fuller explanation will be given in Chapter 8 of the reasons why  $s^2$  is an unbiased estimator of the population variance, whereas  $(SD)^2$  is not.

The square root of  $s^2$  is called the sample standard deviation.

$$s = \sqrt{\frac{\sum x^2}{n-1}} \quad (4-5)$$

Although  $s$ , like  $SD$ , has the advantage of being in the same units of measurement as the original  $X$  values, it is also a biased estimator of the population standard deviation. This bias is much smaller than that for  $SD$  and may be eliminated by use of the correction factor  $a_2$  given in Appendix 9. Thus,  $a_2(s)$  is an unbiased estimator of the population standard deviation.

#### 4.5 TWO MATHEMATICAL PROPERTIES OF THE STANDARD DEVIATION

The following properties hold for both  $SD$  and  $s$ . We will illustrate them by using  $s$ .

1. The numerical value of the sample standard deviation is not changed by the addition of a constant to each  $X$  value. We know that

$$s^2 = \frac{\sum x^2}{n-1} = \frac{\sum (X - \bar{X})^2}{n-1}$$

and if we add a constant  $K$ , which may be negative, to the  $X$  values, the mean becomes

$$\bar{X} = \frac{\sum (X + K)}{n} = \bar{X} + K$$

so that the sample variance of the  $X + K$  values is

$$s_{X+K}^2 = \frac{\sum [(X + K) - (\bar{X} + K)]^2}{n - 1}$$

and

$$s_{X+K} = \sqrt{\frac{\sum (X - \bar{X})^2}{n - 1}}$$

2. Multiplying each  $X$  value by a constant will multiply the sample standard deviation of the  $X$  values by the absolute value of the constant. Thus, if each  $X$  value is multiplied by  $K$ , the mean becomes

$$\bar{KX} = \frac{\sum KX}{n} = K\bar{X}$$

so the sample variance of the  $KX$  values is

$$s_{KX}^2 = \frac{\sum (KX - K\bar{X})^2}{n - 1} = \frac{K^2 \sum (X - \bar{X})^2}{n - 1}$$

and

$$s_{KX} = K \sqrt{\frac{\sum (X - \bar{X})^2}{n - 1}}$$

The student should verify these two results by using the numerical example in Problem 1.

## 4.6 EFFICIENT COMPUTATIONAL METHODS FOR :

**Ungrouped Data.** Use of Eq. (4-5) is unnecessarily laborious because  $\bar{X}$  must be subtracted from each value of  $X$ . To achieve accurate results,  $\bar{X}$  must often be carried to a great many decimal places.<sup>(4)</sup> Thus, the calculation of  $\sum x^2$  requires  $n$  laborious subtractions and the squaring of multidigit numbers. It is usually easier, therefore, to compute variation by subtracting from  $\sum X^2$  a correction term.

$$\sum x^2 = \sum X^2 - \text{correction term}$$

The correction term may be defined in various ways:

$$\text{Correction term} = (\sum X)^2/n = n\bar{X}^2 = \bar{X} \sum X \quad (4-6)$$

In this book we prefer to use<sup>(5)</sup>

$$\sum x^2 = \sum X^2 - \frac{(\sum X)^2}{n} \quad (4-7)$$

<sup>(4)</sup> Although the method of computation described in this section is an efficient one for use with a desk calculator, it will usually not produce an efficient electronic computer program. It seems better to program a computer to subtract  $\bar{X}$  from each  $X$  value.

<sup>(5)</sup> A proof that  $\sum x^2 = \sum X^2 - (\sum X)^2/n$  is as follows:

$$\begin{aligned} \sum x^2 &= \sum (X - \bar{X})^2 = \sum (X^2 - 2\bar{X}X + \bar{X}^2) = \sum X^2 - 2\bar{X} \sum X + n\bar{X}^2 \\ &= \sum X^2 - 2 \frac{(\sum X)^2}{n} + \frac{(\sum X)^2}{n} = \sum X^2 - (\sum X)^2/n \end{aligned}$$

Table 4.4 illustrates the computation procedure for the National Equipment Company data.

**TABLE 4.4: COMPUTATION OF  $\sum X^2$  FOR OBTAINING STANDARD DEVIATION OF NATIONAL EQUIPMENT COMPANY DATA**

( $X$  = tensile strength in pounds; data of Table 4.1)

<i>Item number</i>	$X$	$X^2$
1	130.1	16,926.01
2	132.3	17,503.29
⋮	⋮	⋮
34	180.7	32,652.49
Total	5183.7	795,086.37

Then since

$$s = \sqrt{\frac{\sum X^2 - (\sum X)^2/n}{n - 1}} \quad (4-8)$$

we have for this example

$$s = \sqrt{\frac{795,086.37 - (5183.7)^2/34}{33}} = 12.0 \text{ pounds}$$

The student can verify that the more laborious calculation of  $\sum x^2$ , as given in Table 4.3, yields the same result as the calculation above.

**Grouped Data.** For grouped data one may use the formula

$$s = \sqrt{\frac{\sum fX^2 - (\sum fX)^2/n}{n - 1}} \quad (4-9)$$

where the  $X$  values are the mid-values of the class intervals. This formula is analogous to Eq. (4-8). Table 4.5 illustrates the computation procedure for National Equipment Company data.

**TABLE 4.5: COMPUTATION OF VALUES FOR OBTAINING STANDARD DEVIATION OF NATIONAL EQUIPMENT COMPANY DATA, USING FREQUENCY DISTRIBUTION ( $X$  VALUES ARE CLASS MID-VALUES)**

(Tensile strength in pounds; data of Table 4.1)

<i>Tensile strength</i>	$f$	$X$	$X^2$	$fX$	$fX^2$
125-135	3	130	16,900	390	50,700
135-145	5.5	140	19,600	770	107,800
145-155	12.5	150	22,500	1875	281,250
155-165	9	160	25,600	1440	230,400
165-175	3	170	28,900	510	86,700
175-185	1	180	32,400	180	32,400
Total	34	...	...	5165	789,250



Then, using Eq. (4-9), we have

$$s = \sqrt{\frac{789,250 - (5165)^2/34}{33}} = 11.8 \text{ pounds}$$

## 4.7 RELATIVE DISPERSION

It will be recalled that the data on strength of valve caps made by the National Equipment Company showed  $\bar{X} = 152.5$  pounds and  $s = 12.0$  pounds. Similar test data of 36 valve caps made by a competitor has  $\bar{X} = 147.1$  pounds and  $s = 31.7$  pounds. If we wish to compare the dispersions of these two sets of data, it is *not* incorrect to compare the two  $s$  values. The arithmetic means of the two series are not greatly different, and it is clear that the dispersion of the competitor's test data is greater than that of the National Equipment Company data.

In other instances the comparison takes on a different aspect. The Goodyear Tire and Rubber Company developed a type of cord for use in automobile tires known as "Supertwist"—a cord that not only will stretch more than ordinary cord but also has a longer flex life. The mean flex life, as tested by an apparatus for bending the cord, was 138.64 minutes for Supertwist and 87.66 minutes for regular cord. These tests were made on cord as received from the cotton mill but prior to fabrication in tires. Now, what concerns us at this point of the discussion is the *dispersion* in flex life of the two types of cords. If the two  $s$  values are compared, there seems to be little difference between the two, since the standard deviation for Supertwist is 15.4 minutes, whereas for regular cord it is 14.3 minutes. It must be remembered, however, that the standard deviation of Supertwist is 15.4 minutes in relation to a rather high mean flex life, whereas the standard deviation of regular cord is 14.3 minutes in relation to a rather low mean flex life. Hence we have the concept of relative dispersion  $V$  in which  $s$  is compared to the arithmetic mean. For the sample<sup>(4)</sup>

$$V = \frac{s}{\bar{X}} \quad (4-10)$$

$$\text{For Supertwist: } V = \frac{15.4}{138.64} = 0.1111, \text{ or } 11.1 \text{ percent}$$

$$\text{For regular cord: } V = \frac{14.3}{87.66} = 0.1631, \text{ or } 16.3 \text{ percent}$$

From a comparison of the two  $V$ 's it is apparent that the relative dispersion is much less for Supertwist than for regular cord.

At times it is necessary to compare the dispersions of two series expressed

---

<sup>(4)</sup> An analogous measure may be defined for a population,  $V = \sigma/\mu$ .

in different units. As was noted above, the mean *flex life* of Supertwist was 138.64 minutes, and  $s$  was 15.4 minutes. The mean *tensile strength* of Supertwist was 18.3 pounds, whereas  $s$  was 0.73 pounds. If it is desired to know whether Supertwist shows greater dispersion of tensile strength or of flex life, it is not possible to compare the two  $s$  values, 15.4 minutes and 0.73 pounds. It is absolutely necessary to resort to use of the  $V$ 's. Relative variability in respect to tensile strength is

$$V = \frac{0.73}{18.3} = 0.0399, \text{ or } 4.0 \text{ percent}$$

The  $V$  for flex life was shown above to be 11.1 percent, and it is thus seen that Supertwist is less variable in respect to tensile strength than in respect to flex life.

When dispersions are being compared, three types of situations may be found present, each of which has been illustrated.

1. The series may be expressed in the same units, and the arithmetic means may be the same or nearly the same in size. Here the  $s$  values may validly be compared, and no additional information is obtained by use of the  $V$ 's.

2. The series may be expressed in the same units, but the arithmetic means may be of different size. A comparison of absolute dispersion may be had by considering the  $s$  values, but usually a more meaningful comparison results from comparing relative dispersion through a consideration of the  $V$ 's.

3. The series may be expressed in different units. In this case it is not possible to compare the  $s$  values, but comparison may be made of the  $V$ 's.

## PROBLEMS

1. Given the following  $X$  values: 2, 5, -1, 0, 3
  - a. Calculate  $s$ , using Eq. (4-5).
  - b. Add 3 to each  $X$  value and recalculate  $s$ , using Eq. (4-8).
  - c. Multiply each  $X$  value by 3 and recalculate  $s$ , using Eq. (4-8).
  - d. Show algebraically that the sample standard deviation of the  $n$  values  $I + K(X)$  is  $K$  times the standard deviation of  $X$ , when  $I$  and  $K$  are constants.
2. Show algebraically that  $SD = R/2$  when  $n = 2$ .
3. Show algebraically that all of the following are equivalent:  $(\sum X)^2/n$ ,  $n\bar{X}^2$ ,  $\bar{X} \sum X$ .
4. Calculate  $s$  for the frequency distribution given in Problem 3, Chapter 2.

5. Explain in words the meaning of the following terms:

- a. *Relative dispersion.*
- b. *Sample standard deviation  $s$ .*
- c. *Sample variance  $s^2$ .*
- d. *Variation.*
- e. *Range.*
- f. *Quartile deviation.*
- g. *Mean deviation.*

6. Why can't we use  $\frac{\sum x}{n}$  as a measure of dispersion?

- a. *How does MD overcome this problem?*
- b. *How do  $s$  and SD overcome this problem?*

7. Under what circumstances is  $V$  useful?

8. Show that  $V$  is the sample standard deviation  $s$  of a sample of  $X$  values, each of which has been divided by the arithmetic mean of all the  $X$  values, i.e.,  $X/\bar{X}$ .

9. Correct the following statistics for bias where necessary.

- a.  $R = 10, n = 3$ .
- b.  $SD = 5, n = 5$ .
- c.  $(SD)^2 = 10, n = 9$ .
- d.  $s = 9, n = 4$ .
- e.  $s^2 = 3, n = 100$ .

10. Calculate  $s$  for the competitor's data given in Table 4.1. Compare your answer to that given in Sec. 4.7.

11. If  $\sum x^2 = 4$  and  $n = 2$ , show that  $a_1(SD) = a_2(s)$ .

12. Using the numbers 1, 2, and 3, show that

$$(G)^2 \doteq (\bar{X})^2 - (SD)^2$$

where  $G$  is the geometric mean. Give an algebraic proof that this approximation is exact when  $n = 2$ .

# 5

## Shapes of Frequency Distributions

As was pointed out in Chapter 2, frequency distributions differ in three ways: (1) average value, (2) variability or dispersion, (3) shape. In this chapter we seek to draw together the ideas presented in Chapters 2 through 4. Thus, the relationships between shapes of frequency distributions and averages will be presented, as well as a discussion of how the arithmetic mean and standard deviation of frequency distributions may be used to standardize different frequency distributions with respect to the mean and standard deviation so that they will differ only as to shape.

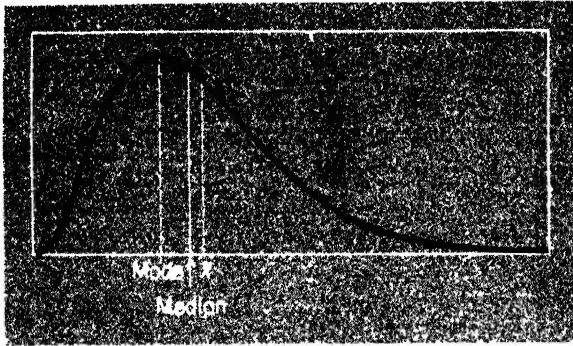
### 5.1 SKEWNESS AND KURTOSIS

A distribution may be defined as skewed if the mean, median, and mode do not all have the same value. From past discussion it is clear that the distribution shown in Chart 5.1 is skewed to the right, or positively, and the distribution shown in Chart 5.2 is skewed to the left, or negatively. Also, recalling that  $\bar{X}$  is akin to the notion of a center of gravity, or balance point, that the median is that value which divides the distribution into equal areas, and that the mode is the value corresponding to the maximum ordinate of the distribution, we should see that

If  $\bar{X} > \text{median} > \text{mode}$ , the skewness is positive (Chart 5.1).

If  $\bar{X} < \text{median} < \text{mode}$ , the skewness is negative (Chart 5.2).

**CHART 5.1: A POSITIVELY SKEWED FREQUENCY CURVE, SHOWING LOCATION OF MEAN, MEDIAN, AND MODE.**



A symmetrical distribution is not skewed, and a skewed distribution is always unsymmetrical. (But an unsymmetrical distribution is not necessarily skewed, for the lack of symmetry can be attributed to various kinds of irregularities.)

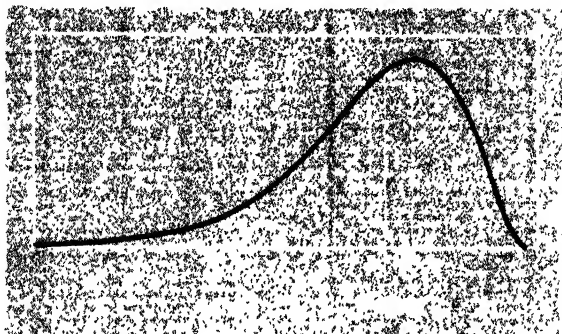
As was pointed out in Chapter 2, a distribution is leptokurtic if it has more small deviations and more large deviations from the mean than normal, but fewer medium-sized ones. It is platykurtic if it has more medium-sized deviations from the mean than normal, but fewer small ones or large ones. In each case the comparison is with a normal curve with the same mean and standard deviation. Charts illustrating kurtosis have been given previously in Sec. 2.7.

One reason that we are interested in kurtosis is that the kurtosis of the population influences the type of average to use for a sample. If a population is very platykurtic, the mid-range is appropriate; if the population is approximately normal, the arithmetic mean should always be used; if the population is very leptokurtic, the median is slightly preferable.

The measurement of kurtosis along with measures of skewness will be discussed in Section 5.3.

## 5.2 STANDARDIZED DISTRIBUTIONS

If we are to make satisfactory comparisons of shapes of frequency distributions, we must eliminate the influence of differences in average value and dispersion from the distributions. The most often used procedure is to cause the arithmetic mean of the distributions to be zero and the standard deviation to be one, or unity. Distributions that have a standard deviation (or, equivalently, variance) of one are said to be *unit distributions*. Distributions with both a standard deviation of one and a mean of zero are said to be standard or *standardized distributions*.

**CHART 5.2: A NEGATIVELY SKEWED FREQUENCY CURVE, SHOWING LOCATION OF MEAN, MEDIAN, AND MODE.**

Consider the sample  $X$  values in Table 5.1. We know that the quantity

$$\Sigma x = \Sigma (X - \bar{X}) = 0$$

Thus, if we transform the  $X$  values into deviations from the mean, the  $x$  values must have a mean of zero, since  $\Sigma x = 0$ . We also know from Sec. 4.5 that this transformation will have no effect on the standard deviation of the  $X$  values.

**TABLE 5.1: CALCULATION OF STANDARDIZED VALUES,  $z$** 

$X$	$x = X - \bar{X}$	$z = x/SD$
8	3	1.5
2	-3	-1.5
6	1	0.5
5	0	0.0
4	-1	-0.5
Sum: 25	0	0
Mean: 5	0	0
SD: 2	2	1

To render the sample standard deviation of a distribution with finite positive variance equal to one, we need only divide each of the values in the distribution by the standard deviation of the distribution. Thus, if we let  $Y = X/(SD_X)$ , we know that

$$\Sigma y^2 = \Sigma \left( \frac{X}{SD_X} - \frac{\bar{X}}{SD_X} \right)^2 = \frac{1}{SD_X^2} \Sigma x^2$$

Therefore

$$SD_Y^2 = \frac{\Sigma y^2}{n} = \frac{1}{SD_X^2} \frac{\Sigma x^2}{n} = \frac{1}{SD_X^2} SD_X^2 = 1$$

A standardized value of a sample is defined as<sup>(1)</sup>

$$z = \frac{X - \bar{X}}{SD} \quad (5-1)$$

In Table 5.1 we see that there is one  $z$  value associated with each value of  $X$  and that the mean of  $z$  is zero and the standard deviation of  $z$  is one.

Standardizing a frequency distribution renders it comparable to other standardized frequency distributions, since the distributions will now differ only with regard to shape, but not with regard to average value or dispersion.<sup>(2)</sup> By means of  $z$  values we may also compare individual items from two different distributions. Suppose that an individual obtains a score of 125 on an achievement test based on volume of output, the mean for all employees being 92 and the standard deviation, 24.5. Then

$$z = \frac{125 - 92}{24.5} = 1.35$$

Suppose also that he is given a score of 84 by his supervisor, this score being based on considerations other than volume of output. Now if for all employees  $\bar{X} = 70$  and  $SD = 21$

$$z = \frac{84 - 70}{21} = 0.67$$

Apparently the person being investigated is more satisfactory with respect to output than with respect to other qualifications. We also see that the  $z$  score may be thought of as a unit of measurement. We may say that the employee in question was given a score by his supervisor which was two-thirds of a standard deviation above average.

### 5.3 MOMENTS

A sample can be described almost completely by the first four moments ( $M_1, m_2, m_3, m_4$ ) and two measures based on the moments ( $a_3$  and

<sup>(1)</sup> Sometimes, especially in educational or psychological testing, a variable or score is "normalized." Although there are various types of normalizing, the most common method is to transform the scores so that their mean is 50 and standard deviation is 10. Also we define the standardized value using  $SD$  rather than  $s$  because  $\bar{X}$  and  $(SD)^2$  are the joint maximum likelihood estimators of  $\mu$  and  $\sigma^2$ . Some texts define  $z$  using  $s$  rather than  $SD$ .

<sup>(2)</sup> In demographic statistics various measures are standardized in a somewhat different manner. For example, death rates in different countries may be standardized with respect to age and sex. The standardized death rate is an estimate of that which would be obtained if the different countries had the same age and sex distribution.

$a_4$ ). By definition

First moment

$$\text{about zero: } M_1 = \frac{\sum X}{n} = \bar{X}, \text{ or } \underline{\text{mean}}$$

Second moment

$$\text{about mean: } m_2 = \frac{\sum x^2}{n} = (SD)^2, \text{ or } \underline{\text{variance}}$$

Third moment

$$\text{about mean: } m_3 = \frac{\sum x^3}{n}, \text{ a } \underline{\text{measure of absolute skewness}}$$

Fourth moment

$$\text{about mean: } m_4 = \frac{\sum x^4}{n}, \text{ a } \underline{\text{measure of absolute kurtosis}}$$

$$a_3 = \frac{m_3}{(SD)^3} = \frac{m_3}{m_2^{3/2}}, \text{ a } \underline{\text{measure of relative skewness}}$$

$$a_4 = \frac{m_4}{(SD)^4} = \frac{m_4}{m_2^2}, \text{ a } \underline{\text{measure of relative kurtosis}}$$

(5-2)

Note that whereas the first moment (denoted by  $M_1$ ) is taken about zero, the second, third, and fourth moments (denoted by lower case  $m$ 's) are taken about the mean. It is easy to see why  $m_3$  measures skewness. When a deviation is cubed, its sign is not changed, but large deviations are increased more, numerically, by cubing them than are small ones. For example,  $4 = 3 + 1$ , but  $4^3 > 3^3 + 1^3$ ;  $64 > 27 + 1$ . On the other hand, we cannot tell whether the distribution is leptokurtic, normal, or platykurtic by looking at  $m_4$ . Nevertheless, it is easy to see why  $m_4$  measures kurtosis. Large deviations are greatly magnified by raising them to the fourth power, and if there is an excess of large deviations the distribution will be leptokurtic. Measuring kurtosis is sometimes called measuring *excess*.

If we were dealing with all of the elements in a population, rather than those in a sample, and if we were to compute measures analogous to those described by Eq. (5-2), we would have what are known as *parameters*. In the same way that *statistics* describe the properties of a sample, *parameters describe the properties of a population*. Generally, Greek letters are used to refer to a parameter. Thus, we have the following moments and measures of relative skewness and kurtosis for a population.

First moment about zero:  $\mu$

Second moment about mean:  $\mu_2$

Third moment about mean:  $\mu_3$

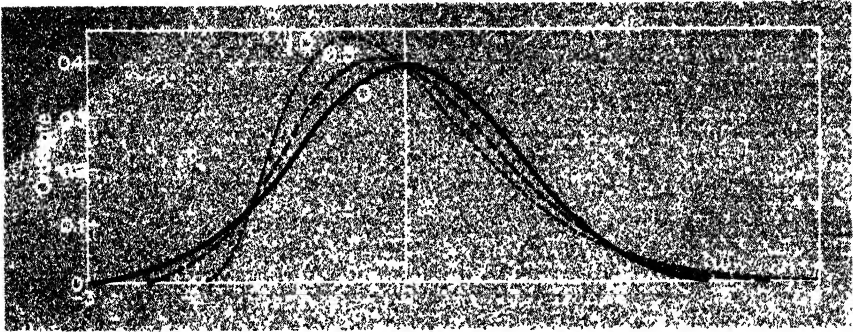
Fourth moment about mean:  $\mu_4$

$$\alpha_3 = \frac{\mu_3}{\sigma^3} = \frac{\mu_3}{\mu_2^{3/2}}$$

$$\alpha_4 = \frac{\mu_4}{\sigma^4} = \frac{\mu_4}{\mu_2^2}$$



**CHART 5.3: PEARSONIAN TYPE III CURVES FOR VARYING DEGREES OF SKEWNESS:**  
**NESS:**  $\alpha_3 = 0$ ;  $\alpha_3 = 0.5$ ;  $\alpha_3 = 1.0$ .



Source: Dudley J. Cowden, *Statistical Methods in Quality Control* (Englewood Cliffs, N.J.: Prentice-Hall Inc., 1957), Fig. 2.7, p. 19.

The second moment about the mean,  $\mu_2$ , is defined as the population variance,  $\sigma^2$ , for an infinitely large population.

For a normal distribution  $\alpha_3 = 0$ . A good idea of the relationship between the magnitude of  $\alpha_3$  and the degree of skewness can be obtained by examining Chart 5.3. It appears that if the absolute value of  $\alpha_3$  is greater than 0.5, there is considerable skewness. For a normal distribution  $\alpha_4 = 3$ ; for a leptokurtic distribution  $\alpha_4 > 3$ ; for a platykurtic distribution  $\alpha_4 < 3$ . Because of these relationships,  $\alpha_4 - 3$  is sometimes taken as a measure of kurtosis.<sup>(3)</sup> A good idea of the relationship between the value of  $\alpha_4 - 3$  and the degree of kurtosis can be obtained by examining Chart 5.4. For a rectangular distribution  $\alpha_4 = 1.8$  and  $\alpha_4 - 3 = -1.2$ .

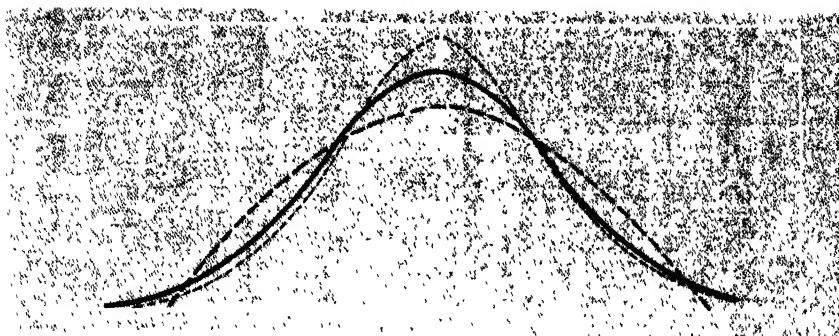
**Computation.** It is evident that the second through fourth sample moments given in Eq. (5-2) could be calculated by forming  $n$  values for  $x = (X - \bar{X})$  and finding  $\sum x^2$ ,  $\sum x^3$ , and  $\sum x^4$ . Practically, this method is not usually followed because of its laborious nature. It is easier first to compute the moments about zero,  $M_r$ .

$$\left. \begin{aligned} M_1 &= \frac{\sum X}{n} \\ M_2 &= \frac{\sum X^2}{n} \\ M_3 &= \frac{\sum X^3}{n} \\ M_4 &= \frac{\sum X^4}{n} \end{aligned} \right\} \quad (5-3)$$

<sup>(3)</sup> Alternate systems of notation are sometimes used:

$$\begin{aligned} \gamma_1 &= \alpha_3 \quad \text{and} \quad \gamma_2 = \alpha_4 - 3 \\ \sqrt{\beta_1} &= \alpha_3 \quad \text{and} \quad \beta_2 = \alpha_4 \end{aligned}$$

**CHART 5.4: PEARSONIAN TYPE II CURVES WITH  $\alpha_3 = 0$  FOR VARYING DEGREES OF KURTOSIS:  $\alpha_4 = 1.8$ ;  $\alpha_4 = 2.2$ ;  $\alpha_4 = 3$ ;  $\alpha_4 = 5$ .**



Source: Dudley J. Cowden, *Statistical Methods in Quality Control* (Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1957), Fig. 2.8, p. 21.

From these we compute the moments about the mean.

$$\left. \begin{aligned} m_2 &= M_2 - M_1^2 \\ m_3 &= M_3 - 3M_2M_1 + 2M_1^3 \\ m_4 &= M_4 - 4M_3M_1 + 6M_2M_1^2 - 3M_1^4 \end{aligned} \right\} \quad (5-4)$$

Computation of the  $M_r$  values is given in Table 5.2.

We now proceed to compute the moments about the mean,  $a_3$  and  $a_4$ .

$$m_2 = 94.7 - (9.7)^2 = 0.61$$

$$m_3 = 930.1 - 3(94.7)(9.7) + 2(9.7)^3 = -0.324$$

$$m_4 = 9185.9 - 4(930.1)(9.7) + 6(94.7)(9.7)^2 - 3(9.7)^4 = 1.1737$$

$$a_3 = \frac{-0.324}{0.4763} = -0.68$$

$$a_4 = \frac{1.1737}{0.3721} = 3.15$$

**TABLE 5.2: COMPUTATION OF MOMENTS ABOUT ZERO**

$X$	$X^2$	$X^3$	$X^4$
8	64	512	4096
9	81	729	6561
9	81	729	6561
10	100	1000	10,000
10	100	1000	10,000
10	100	1000	10,000
10	100	1000	10,000
10	100	1000	10,000
10	100	1000	10,000
10	100	1000	10,000
11	121	1331	14,641
Sum: 97	947	9301	91,859
$M_r$ : 9.7	94.7	930.1	9185.9

Thus the distribution has considerable negative skewness, but it is only slightly leptokurtic.<sup>(4)</sup>

Sample moments can be calculated from a frequency distribution. Calculation of  $M_1 = \bar{X}$  and  $m_2 = (SD)^2$  have already been illustrated. Various texts illustrate the calculation of higher sample moments from a frequency distribution.<sup>(5)</sup>

## 5.4 FISHER'S $k$ -STATISTICS

We noted in the last chapter that  $(SD)^2 = m_2$  is a biased estimator of  $\sigma^2$ , the population variance. It was also stated that this bias could be removed by multiplying  $(SD)^2$  by  $n/(n-1)$ . In a similar manner it is true that  $m_3$  and  $m_4$  are biased estimators of  $\mu_3$  and  $\mu_4$ . Therefore, R. A. Fisher suggested the following scheme for describing a sample:

$$\left. \begin{aligned} k_1 &= M_1 = \bar{X} \\ k_2 &= \frac{n}{n-1} m_2 = s^2 \\ k_3 &= \frac{n^2}{(n-1)(n-2)} m_3 \\ k_4 &= \frac{n^2}{(n-1)(n-2)(n-3)} [(n+1)m_4 - 3(n-1)m_2^2] \\ g_1 &= \frac{k_3}{k_2^{3/2}} \\ g_2 &= \frac{k_4}{k_2^2} \end{aligned} \right\} \quad (5-5)$$

Being unbiased estimators,  $g_1$  and  $g_2$  give a more accurate idea of the skewness and kurtosis in the population than do  $a_3$  and  $a_4$ .<sup>(6)</sup> Also,  $g_1$  is analogous to  $a_3$  and  $g_2$  to  $a_4 - 3$ . If we call  $\gamma_1$  and  $\gamma_2$  the parameters estimated by  $g_1$  and  $g_2$ , both  $\gamma_1$  and  $\gamma_2$  are zero for a normal population. As the sample size increases, the difference between  $a_3$  and  $g_1$  becomes smaller, as does the difference between  $a_4 - 3$  and  $g_2$ .

<sup>(4)</sup> For large samples, the significance of  $a_3$  and  $a_4$  may be tested by comparing  $a_3$  and  $a_4 - 3$  with their standard errors, which are given approximately by

$$\sigma_{a_3} \doteq \sqrt{6/n} \quad \sigma_{a_4} \doteq \sqrt{24/n}$$

Testing for significance of skewness and kurtosis is explained in the appendix to Chapter 7.

<sup>(5)</sup> For example, F. E. Croxton, D. J. Cowden, and S. Klein, *Applied General Statistics*, 3rd ed. (Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1967), Chapter 10.

<sup>(6)</sup> The significance of  $g_1$  and  $g_2$  can be tested by comparing them with their standard errors. For all but very small samples,  $g_1$  and  $g_2$  are distributed almost normally for normal populations, with standard errors

$$\sigma_{g_1} = \sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}} \quad \sigma_{g_2} = \sqrt{\frac{24n(n-1)^3}{(n-3)(n-2)(n+3)(n+5)}}$$

## PROBLEMS

1. Karl Pearson has suggested a measure of skewness

$$Sk = \frac{3(\bar{X} - Med)}{SD}$$

which has theoretical maximum and minimum values of  $\pm 3$ . Explain why this statistic measures skewness.

2. Standardize the following  $X$  values

1, 2, 3, 2, 5, 0

- Calculate the standard deviation of the standardized values.
- Calculate the mean of the standardized values.

3. Given the following  $X$  values

3, 3, 5, 5, 5, 10, 10, 15

- Calculate the first moment about zero and the second through fourth moments about the mean.
- Calculate a measure of relative skewness.
- Calculate a measure of relative kurtosis.
- Characterize in words the shape of the distribution.

4. Using the data in problem 3 above, calculate Fisher's  $k_1$  through  $k_4$ ;  $g_1$  and  $g_2$ . Repeat part d. of Problem 3. What advantage is obtained by use of  $k$ -statistics over moments?

- What is the relative dispersion,  $V$ , of a standardized variable?
- What is the range of a symmetrical standardized variable whose greatest positive value is  $Kz$ , where  $K$  is a positive finite number?

6. Why is  $M_1$  called a moment about zero while  $m_2$  is called a moment about the mean?

7. Show algebraically that

$$m_4 = M_4 - 4M_3M_1 + 6M_2M_1^2 - 3M_1^4$$

and develop a computation formula for  $m_5$ .

8. R. C. Geary has suggested a measure of kurtosis

$$Kr = \frac{\sum |X - \bar{X}|}{\sqrt{n} \sum (X - \bar{X})^2}$$

Explain why this statistic will be larger for a platykurtic distribution than for a leptokurtic distribution.

# 6

## Probability and Some Discrete Probability Distributions

Statistics is concerned, among other things, with the making of decisions by the application of the theory of probability to observed numerical data. Consider this problem. If you are willing to tolerate only one percent of defective units in a lot of goods, would you accept or reject the lot if a random sample of 50 units from a very large lot contains seven defective items? It can be shown that the probability of having seven or more defectives in a sample of 50 from an infinite population that is one percent defective is 0.000001. What would be your decision?

Some of the most advanced and subtle methods of mathematics are used in probability, but we can develop the basic ideas without recourse to anything more advanced than high school algebra.

Probabilistic reasoning is sometimes used when deterministic reasoning cannot grapple with a problem. For example, if we know the total assets of a firm and the total liabilities, we can easily calculate the total value of the proprietary section of the balance sheet. This is deterministic reasoning. On the other hand, if we see an honest man about to throw a pair of honest dice, we do not know what total will turn up on the dice, but probabilistic reasoning can be applied to find the most probable total, i.e., the total that will turn up most often. Probabilistic reasoning is used in such various fields as gambling, insurance, theoretical physics, biology, economics, among others.

## 6.1 DEFINITIONS OF PROBABILITY

**The Classical Definition.** To make the idea of probability concrete, consider an experiment in which an event can occur in a certain number of ways, say  $r$  different ways, and can fail to occur in a certain number, say  $s$ , different ways. If  $A$  is a success and  $B$  is a failure, the probability of a success is

$$\text{Prob}(A) = \frac{r}{r + s}$$

This is, the *probability of an event is the ratio of the number of ways an event can occur to the total number of possible outcomes, when each possible outcome is equally likely.*

Similarly, the probability of a failure is

$$\text{Prob}(B) = \frac{s}{r + s}$$

or the ratio of the number of ways the event can fail to occur to the total number of possible outcomes.

Notice that  $\text{Prob}(A)$  and  $\text{Prob}(B)$  are both fractions because  $r$  is equal to or smaller than  $r + s$ , as is  $s$ . Notice also that

$$\text{Prob}(A) + \text{Prob}(B) = \frac{r}{r + s} + \frac{s}{r + s} = 1$$

That is, the probability of success plus the probability of failure equals one.

As an example, consider the tossing of an unbiased coin. Suppose we agree that a head is a success and a tail is a failure. Then a success can occur in only one way; that is, by a head showing on a toss, and a failure can occur in only one way; that is, by a tail showing on a toss. Hence the total number of possible outcomes is two (head or tail). Then for this **example**

$$\begin{aligned} r &= 1 & s &= 1 \\ \text{Prob}(A) &= \frac{r}{r + s} = \frac{1}{1 + 1} = \frac{1}{2} \end{aligned}$$

$$\text{Prob}(B) = \frac{s}{r + s} = \frac{1}{1 + 1} = \frac{1}{2}$$

$$\text{Prob}(A) + \text{Prob}(B) = \frac{1}{2} + \frac{1}{2} = 1$$

As another example, consider the matching game. You win if a match occurs when two coins are tossed, and your opponent wins if a match does not occur. Is the game fair? The possible outcomes of the game are

TT TH HT HH

Hence there are four possible outcomes. You win if TT or HH show. You lose if TH or HT shows. Consequently,

$$r = 2 \quad s = 2 \quad r + s = 4$$

$$\text{Prob } (A) = \frac{2}{4} = \frac{1}{2}$$

$$\text{Prob } (B) = \frac{2}{4} = \frac{1}{2}$$

$$\text{Prob } (A) + \text{Prob } (B) = 1$$

The game is obviously fair if the coins are fair.

These are among the most elementary examples of the use of probability, and when problems become slightly complicated, finding the solution may become greatly complicated. For example, what is the probability that the player who casts the dice in a crap game will win?<sup>(1)</sup>

**Another Definition of Probability.** We tentatively defined probability as the ratio of the number of ways an event can occur to the total number of possible outcomes, when each possible outcome is equally likely. This definition is not completely satisfactory because it involves circular reasoning, for "equally likely" is only another way of saying "equally probable." Thus, the previous definition of probability used the idea of probability as a part of the definition. Also, if we are playing with loaded dice, each possible outcome is not equally likely.

In recent years intensive effort has been expended to give a precise mathematical meaning to the term probability. Many philosophers and mathematicians have offered the view of probabilities as long-run relative frequencies. Thus, another interpretation of the meaning of probability is given by stating that *probability is the limit of the relative frequency of successes in an infinite sequence of trials*. This definition also involves logical difficulties, for we cannot be sure that a limit exists. But the definition permits us to estimate a probability experimentally.<sup>(2)</sup> For example, if an urn contains a very large number of balls, some of which are red and the others white, and if we draw a sample of  $n$  balls from the urn after thorough mixing, always replacing each ball before the next is drawn, we should expect the ratio of red balls to total balls in the sample—the relative frequency of the red balls—to be almost the proportion in the urn if  $n$  is large enough. As  $n$  approaches infinity,  $p$  approaches  $P$ , where  $p$  is the proportion of red balls in the sample, and  $P$  is the proportion of red balls in the urn population. Probability can also be

<sup>(1)</sup> It is approximately 0.493. See Paul Peach, *An Introduction to Industrial Statistics and Quality Control*, 2nd ed. (Raleigh, N.C.: Edwards & Broughton, 1947), p. 6.

<sup>(2)</sup> Quoting from S. S. Wilks, *Elementary Statistical Analysis* (Princeton, N.J.: Princeton University Press, 1948), p. 61 (with slight change in symbols):

"If (1) whenever a series of many trials is made, the ratio of the number of times  $A$  occurred to the total number of trials is nearly  $P$ , and if (2) the ratio is nearer to  $P$  when longer series of trials are made, then we agree in advance to define the probability of  $A$  as  $P$ , or more briefly  $\text{Prob } (A) = P$ ."

estimated experimentally under our second definition when the different outcomes are not equally likely. In this case the first definition of probability is not applicable. For example, we can estimate the probability of obtaining a one-spot from a six-sided die when the die is loaded in a manner unknown to us.

✓ **The Subjective Approach.** The previous definition of probability involves two basic practical difficulties. First, since an infinite sequence of trials is required to determine the probability of an event with exactitude, such a determination is impossible as a practical matter. Second, businessmen and others often deal with unique events, i.e., events that have never happened before and will never happen again under precisely the same conditions. Thus, theorists such as J. M. Keynes and L. J. Savage are among those who in recent years have advocated viewing probability as a primitive intuitive, or personal concept. The probability of an event might, therefore, be thought of as a degree of rational belief. This view of probability is called *subjective* or *personalistic*.

Under the objective or relative frequency doctrine a statement such as: "The probability that my business will fail this year is 0.02" has no meaning. Under the subjective doctrine the statement has legitimate meaning and may be utilized in decision making in its own right or as a supplement to quasi-objective probabilities.

A criticism of subjective probability that is sometimes voiced is that, given a single event, different people will assign different probabilities to the event. These differences, of course, reflect the varying backgrounds of the persons involved. Many statisticians think that the probability assigned to an event should follow from the results of the trials of the event and not depend on matters such as personal prejudices. However, it is clear that the long-run frequency approach involves subjective elements, since judgment is involved in the application of the definition.

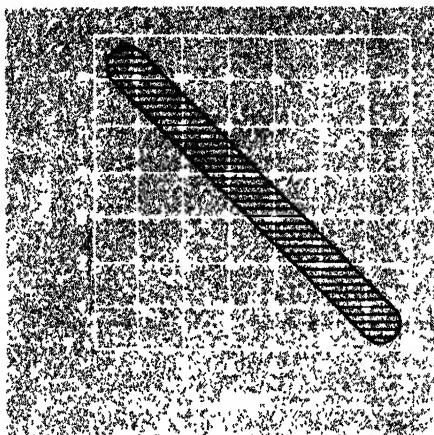
✓ It should be clear that no single definition of the term "probability" is completely satisfactory. In some cases the objective approach is applicable, and in others the subjective approach is more appropriate. We will make use of both approaches in this text.

## 6.2 EVENTS AND SAMPLE SPACES

In the language of probability, an *elementary*, or *simple*, event is an outcome of an experiment that cannot be decomposed into a combination of other elementary events according to the probability model under consideration. Thus in the experiment of tossing a coin, one elementary event is "heads" and the other "tails." If an event can be decomposed into elementary events, the event is called a *compound* or *composite event*. The event



**CHART 6.1: SAMPLE SPACE FOR EXPERIMENT OF TOSSING TWO IDENTICAL SIX-SIDED DICE.**



“heads” in a coin-tossing experiment is an elementary event, but the event “heads or tails” is a compound event, since it can logically be decomposed into two elementary events. If two elementary events cannot both happen on the same trial, they are said to be *mutually exclusive*, and if the outcome of one event does not influence the outcome of another event, the two events are said to be *independent*.

The collection of all possible elementary events associated with a given experiment forms what is known as a *sample space*. Chart 6.1 shows a sample space associated with the experiment of tossing two identical six-sided dice. We notice from Chart 6.1 that the sample space consists of 36 *sample points* which correspond to each possible elementary event associated with this experiment. Using Chart 6.1, one can easily see that the probability of throwing a total of 7 spots is  $\frac{6}{36}$ , or  $\frac{1}{6}$ . The sample points corresponding to the *compound event* of throwing a total of 7 spots is enclosed in the cross-hatched region of Chart 6.1.

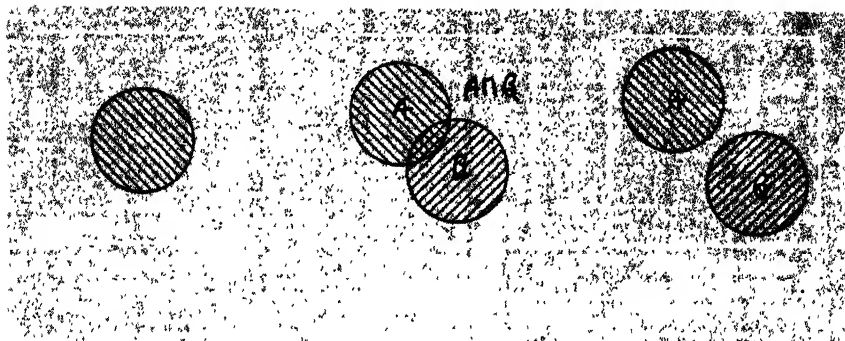
In a general but more abstract manner we may denote a sample space by means of a box and the events of interest by figures such as circles within the box. The left diagram in Chart 6.2 is such a representation and is called a *Venn diagram*. The two right diagrams are also examples of Venn diagrams and will be discussed in later sections.

### 6.3 PROBABILITY AXIOMS

The following axioms are necessary for the development of probability theory. Given an experiment:

1. Each elementary event, or combination of elementary events, must have associated with it a probability greater than or equal to zero but less

CHART 6.2: USE OF VENN DIAGRAMS.



than or equal to one. Thus, if  $A$  is an event within a sample space,

$$0 \leq \text{Prob}(A) \leq 1 \quad (6-1)$$

2. The probability of an entire sample space is one. Thus, if  $S$  represents an entire sample space,

$$\text{Prob}(S) = 1 \quad (6-2)$$

3. The probability that one or the other or both of two *mutually exclusive* events will occur is equal to the sum of the individual probabilities of these events. Thus

$$\text{Prob}(A \cup B) = \text{Prob}(A) + \text{Prob}(B) \quad (6-3)$$

when  $A$  and  $B$  are mutually exclusive events. Notice that the symbol " $\cup$ " (cup) is read "or" and is used in the inclusive sense to mean the set of points belonging to  $A$  or to  $B$  or both. Formally, it is the union of the sets  $A$  and  $B$ . (See the right diagram in Chart 6.2.)

## 6.4 PROBABILITY AND INDEPENDENT EVENTS

The condition of statistical independence is defined by the relationship

$$\text{Prob}(A | B) = \text{Prob}(A) \quad (6-4)$$

and if Eq. (6-4) holds, then the following holds as well:

$$\text{Prob}(B | A) = \text{Prob}(B) \quad (6-5)$$

The symbol " $|$ " is read "given." Thus, if the event  $A$  is unaffected by the fact that  $B$  has occurred,  $A$  and  $B$  are independent events and Eq. (6-4) holds by definition. The probability that heads will be thrown on the second toss of a fair coin, given that heads appeared on the first toss, is

$$\text{Prob}(\text{Head}_2 | \text{Head}_1) = \text{Prob}(\text{Head}_2) = \frac{1}{2}$$

Successive tosses of a fair coin are considered to be statistically independent events.

If two events are independent, the probability that they will both happen is

$$\text{Prob}(A \cap B) = \text{Prob}(A) \cdot \text{Prob}(B) \quad (6-6)$$

where " $\cap$ " (cap) is read "and." It is the intersection of the sets  $A$  and  $B$ . (See the center diagram in Chart 6.2.) Equation (6-6) is known as the multiplication rule for two independent events. With two fair coins, the probability that two heads will be tossed on the same trial is

$$\begin{aligned} \text{Prob}(\text{Head}_1 \cap \text{Head}_2) &= \text{Prob}(\text{Head}_1) \cdot \text{Prob}(\text{Head}_2) \\ &= \left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = \frac{1}{4} \end{aligned}$$

Putting these ideas together, let us calculate the probability of tossing a total of 7 spots with a pair of identical fair dice, which we already know to be  $\frac{1}{6}$ .

$$\text{Prob}(1 \cap 6 \cup 2 \cap 5 \cup 3 \cap 4 \cup 4 \cap 3 \cup 5 \cap 2 \cup 6 \cap 1) =$$

by Eq. (6-3)

$$\text{Prob}(1 \cap 6) + \text{Prob}(2 \cap 5) + \cdots + \text{Prob}(6 \cap 1) =$$

by Eq. (6-6)

$$\begin{aligned} &\text{Prob}(1) \cdot \text{Prob}(6) + \text{Prob}(2) \cdot \text{Prob}(5) + \cdots + \text{Prob}(6) \cdot \text{Prob}(1) \\ &= \left(\frac{1}{6}\right)\left(\frac{1}{6}\right) + \left(\frac{1}{6}\right)\left(\frac{1}{6}\right) + \cdots + \left(\frac{1}{6}\right)\left(\frac{1}{6}\right) = \frac{6}{36} = \frac{1}{6} \end{aligned}$$

## 6.5 PROBABILITY AND DEPENDENT EVENTS

Consider the following collection of men and women found in a large office.

<i>Color of hair</i>	<i>Men</i>	<i>Women</i>	<i>Total</i>
Black	10	90	100
Red	50	50	100
Total	60	140	200

The probability of drawing a man at random from this collection is<sup>(3)</sup>

$$\text{Prob}(\text{man}) = \frac{60}{200} = 0.3$$

The probability of drawing a red-headed person at random from this collection is

$$\text{Prob}(\text{red}) = \frac{100}{200} = 0.5$$

What is the probability of drawing a red-headed person from this collection

<sup>(3)</sup> The meaning of the term "at random" will be treated in greater detail in a later chapter. The essence of the term is that the sample is taken in such a manner that every item and every collection of items in the population being sampled have an equal chance of being selected.

given that he is a man, Prob (red | man)? In this problem, the fact that we have specified that the individual in question is a man has *reduced* the sample space from 200 persons to the 60 persons who are men. Thus, using 60 sample points, we calculate that

$$\text{Prob (red | man)} = \frac{\text{Prob (man} \cap \text{red)}}{\text{Prob (man)}} = \frac{50/200}{60/200} = \frac{50}{60} = \frac{5}{6}$$

The division by  $\frac{60}{200}$  has effectively reduced the sample space from 200 sample points to 60 sample points. Generalizing this result, we have

$$\text{Prob (B | A)} = \frac{\text{Prob (A} \cap \text{B)}}{\text{Prob (A)}} \quad (6-7)$$

assuming that Prob (A) is not equal to zero.

To find Prob (A  $\cap$  B) under conditions of statistical dependence we need only rearrange Eq. (6-7) to form

$$\begin{aligned} \text{Prob (A} \cap \text{B)} &= \text{Prob (A)} \cdot \text{Prob (B | A)} \\ &= \text{Prob (B)} \cdot \text{Prob (A | B)} \end{aligned} \quad (6-8)$$

The probability of drawing a person who is both red-headed and male from this collection is

$$\begin{aligned} \text{Prob (man} \cap \text{red)} &= \text{Prob (man)} \cdot \text{Prob (red | man)} \\ &= (\frac{60}{200})(\frac{5}{6}) = \frac{50}{200} = 0.25 \end{aligned}$$

All four of the probabilities calculated according to Eq. (6-8) are given in the body of the table below. The probabilities are referred to as *joint probabilities*. Notice that the joint probabilities sum to the *marginal probabilities*, which are calculated by dividing each of the row and column totals in the previous table by the grand total. These marginal probabilities are given below as row and column totals.

<i>Color of hair</i>	<i>Men</i>	<i>Women</i>	<i>Total</i>
Black	0.05	0.45	0.50
Red	0.25	0.25	0.50
Total	0.30	0.70	1.00 ,

## 6.6 GENERALIZATION AND EXTENSIONS OF THE FORMULAS

In Sec. 6.3 through 6.5 we have, in reality, discussed special cases of two formulas. The first formula, which is a generalization of Eq. (6-3), is called the *generalized addition rule for two events*.

$$\text{Prob (A} \cup \text{B)} = \text{Prob (A)} + \text{Prob (B)} - \text{Prob (A} \cap \text{B)} \quad (6-9)$$

Referring to the Venn diagrams given in Chart 6.2, we see that when A and B are mutually exclusive events (shown in the right diagram), the term Prob (A  $\cap$  B) vanishes from Eq. (6-9), since A and B have no points in

common. However, when  $A$  and  $B$  are not mutually exclusive (shown in the middle diagram of Chart 6.2), the simple addition rule of Eq. (6-3) can no longer hold, since it would involve double counting of the space occupied by  $A$  and  $B$  jointly. The intersection ( $A \cap B$ ) must be subtracted to avoid this double counting.

For example, to find the probability that a person drawn will be a woman or have black hair, we evaluate

$$\begin{aligned}\text{Prob}(\text{woman} \cup \text{black}) &= \text{Prob}(\text{woman}) + \text{Prob}(\text{black}) - \text{Prob}(\text{woman} \cap \text{black}) \\ &= 0.70 + 0.50 - 0.45 = 0.75\end{aligned}$$

Similarly, we evaluate the only other sampling alternative, namely, the probability of drawing a person who is a man with red hair.

$$\text{Prob}(\text{man} \cap \text{red}) = 0.25$$

The total of the two probabilities is one, indicating that the sample space has been exhausted. The generalized addition rule for three events is

$$\begin{aligned}\text{Prob}(A \cup B \cup C) &= \text{Prob}(A) + \text{Prob}(B) + \text{Prob}(C) \\ &\quad - \text{Prob}(A \cap B) - \text{Prob}(A \cap C) \\ &\quad - \text{Prob}(B \cap C) + \text{Prob}(A \cap B \cap C) \quad (6-10)\end{aligned}$$

and for four events is

$$\begin{aligned}\text{Prob}(A \cup B \cup C \cup D) &= \text{Prob}(A) + \text{Prob}(B) + \text{Prob}(C) + \text{Prob}(D) \\ &\quad - \text{Prob}(A \cap B) - \text{Prob}(A \cap C) - \text{Prob}(A \cap D) \\ &\quad - \text{Prob}(B \cap C) - \text{Prob}(B \cap D) - \text{Prob}(C \cap D) \\ &\quad + \text{Prob}(A \cap B \cap C) + \text{Prob}(A \cap B \cap D) \\ &\quad + \text{Prob}(A \cap C \cap D) + \text{Prob}(B \cap C \cap D) \\ &\quad - \text{Prob}(A \cap B \cap C \cap D) \quad (6-11)\end{aligned}$$

and so on for larger numbers of events.

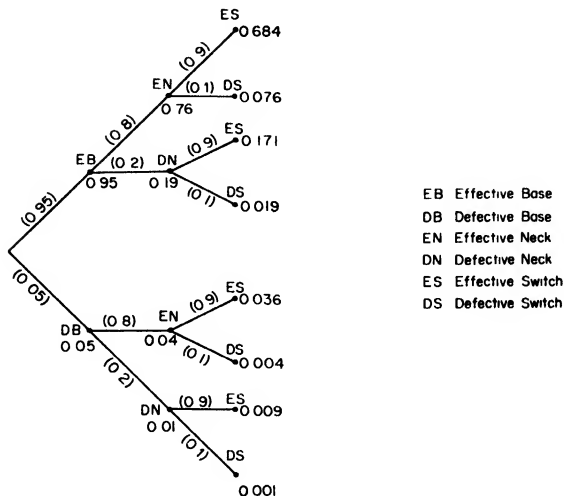
The second major formula that we have discussed is called the *generalized multiplication rule*. For two events it is given by

$$\text{Prob}(A \cap B) = \text{Prob}(A) \cdot \text{Prob}(B | A)$$

We noted that Eq. (6-7) could be derived from this formula and that under conditions of statistical independence, Eq. (6-7) reduced to Eq. (6-5). To illustrate further the use of the formula, consider the following problem: An assembly room for a lamp manufacturer receives lamp components, some of which are defective. The components received on a given day together with the percentage of each component group which was defective are given below.

Components	% Defective	% Effective
Base	5	95
Neck	20	80
Switch	10	90

If three necessary components are selected at random from this lot of components, what is the probability that the assembled lamp will have no defective



parts? The solution to this problem may be couched in terms of a *probability* tree such as the one given above. To achieve a lamp with no defective parts we must follow the upper branch. At each intersection of branches, called a *node*, the probability of reaching that node is given and is derived by use of the multiplication rule. Thus

$$\begin{aligned}\text{Prob}(\text{effective base} \cap \text{effective neck}) &= \text{Prob}(\text{effective base}) \cdot \text{Prob}(\text{effective neck}) \\ &= (0.95)(0.80) = 0.76\end{aligned}$$

And so on for other nodes. The simple marginal probabilities of each branch of the tree are also given. It is interesting to note that the probabilities for corresponding nodes sum vertically to unity since they represent mutually exclusive events, which exhaust the sample space.

$$\begin{aligned}\text{Prob}(\text{effective base} \cup \text{defective base}) &= \text{Prob}(\text{effective base}) \\ &\quad + \text{Prob}(\text{defective base}) \\ &= 0.95 + 0.05 = 1\end{aligned}$$

Finally, the student should verify that the probability tree implies that a generalized multiplication rule for  $K$  events may be written

$$\begin{aligned}\text{Prob}(A_1 \cap A_2 \cap \cdots \cap A_K) &= \text{Prob}(A_1) \text{Prob}(A_2 | A_1) \\ &\quad \cdot \text{Prob}(A_3 | A_1 \cap A_2) \cdots \\ &\quad \cdot \text{Prob}(A_K | A_1 \cap A_2 \cap \cdots \cap A_{K-1}) \quad (6-12)\end{aligned}$$

In the case where the events are independent, Eq. (6-12) reduces to

$$\text{Prob}(A_1 \cap A_2 \cap \cdots \cap A_K) = \text{Prob}(A_1) \cdot \text{Prob}(A_2) \cdots \text{Prob}(A_K) \quad (6-13)$$

which is the case illustrated by the probability tree. Thus the probability that the lamp will have no defective parts is  $(0.95)(0.80)(0.90) = 0.684$ .

## 6.7 DISCRETE PROBABILITY DISTRIBUTIONS

A probability distribution is a rule that assigns a probability to every possible outcome of an experiment. An event whose numerical value is determined by the outcome of an experiment is called a *variate* or often a *random variable*.

Given the assumptions set out earlier in this chapter, we say that if an experiment has only a finite number of outcomes, it possesses a *discrete* probability distribution. Tables 6.1 and 6.2 give numerical values for two discrete probability distributions associated with examples previously discussed in this chapter. These two distributions are graphically presented in Charts 6.3 and 6.4. The student should trace the development of these distributions from the earlier examples.

Associated with almost any probability distribution is an arithmetic mean and variance. The arithmetic mean of a probability distribution is called the *expected value* or *expectation* of the distribution. If  $d$  is a discrete random variable, the expected value of  $d$ ,  $E(d)$ , is calculated by weighting the possible values of  $d$  by the probabilities associated with these values and summing. Since the probabilities sum to one, we are using a formula equivalent to the one for the arithmetic mean given by Eq. (3-3). Thus

$$E(d) = \sum [d \cdot \text{Prob}(d)] \quad (6-14)$$

The expected value of the probability distribution associated with the experiment of tossing two identical fair dice is

$$\begin{aligned} E(d) &= 2\left(\frac{1}{36}\right) + 3\left(\frac{2}{36}\right) + \cdots + 12\left(\frac{1}{36}\right) \\ &= \frac{252}{36} = 7 \end{aligned}$$

The variance of a discrete probability distribution is defined as

$$\sigma_d^2 = E[d - E(d)]^2 = \sum \{[d - E(d)]^2 \cdot \text{Prob}(d)\} \quad (6-15)$$

Thus, for the dice experiment

$$\begin{aligned} \sigma_d^2 &= (-5)^2\left(\frac{1}{36}\right) + (-4)^2\left(\frac{2}{36}\right) + \cdots + (5)^2\left(\frac{1}{36}\right) \\ &= 5.8333 \cdots \end{aligned}$$

In Tables 6.1 and 6.2 we have also cumulated the two probability distributions. Examples of the use of a cumulative probability distribution are as follows. For the dice experiment

$$\text{Prob}(d < 5 \text{ spots}) = \frac{10}{36}$$

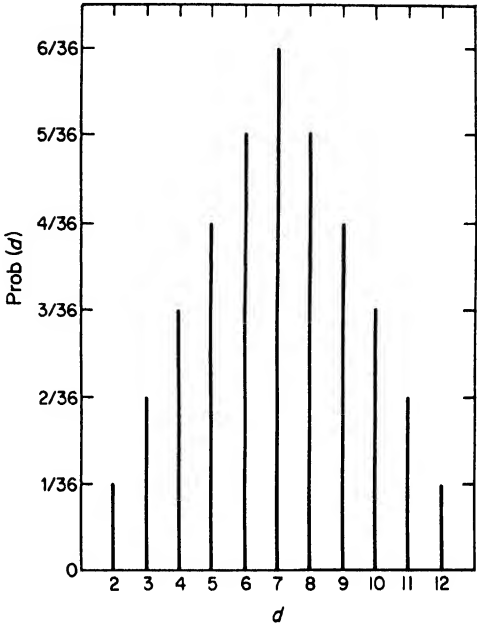
and for the lamp assembly experiment

$$\text{Prob}(d < 1 \text{ defective item}) = 0.967$$

**TABLE 6.1: PROBABILITY DISTRIBUTION ASSOCIATED WITH TOSSING TWO FAIR DICE**

Number of spots $d$	Prob ( $d$ )	Cumulative Prob ( $d$ )
2	$\frac{1}{36}$	$\frac{1}{36}$
3	$\frac{2}{36}$	$\frac{3}{36}$
4	$\frac{3}{36}$	$\frac{6}{36}$
5	$\frac{4}{36}$	$\frac{10}{36}$
6	$\frac{5}{36}$	$\frac{15}{36}$
7	$\frac{6}{36}$	$\frac{21}{36}$
8	$\frac{5}{36}$	$\frac{26}{36}$
9	$\frac{4}{36}$	$\frac{30}{36}$
10	$\frac{3}{36}$	$\frac{33}{36}$
11	$\frac{2}{36}$	$\frac{35}{36}$
12	$\frac{1}{36}$	$\frac{36}{36}$
Total	1.0	...

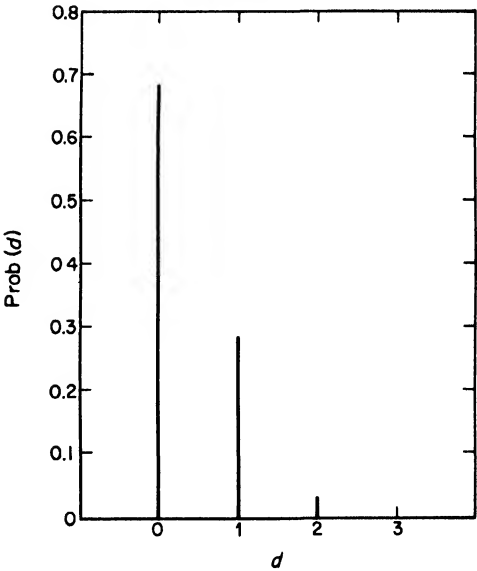
**CHART 6.3: GRAPH OF PROBABILITY DISTRIBUTION GIVEN IN TABLE 6.1.**



**TABLE 6.2: PROBABILITY DISTRIBUTION ASSOCIATED WITH LAMP ASSEMBLY EXPERIMENT**

Number of defective parts $d$	Prob ( $d$ )	Cumulative Prob ( $d$ )
0	0.684	0.684
1	0.283	0.967
2	0.032	0.999
3	0.001	1.000
Total	1.0	...

**CHART 6.4: GRAPH OF PROBABILITY DISTRIBUTION GIVEN IN TABLE 6.2.**

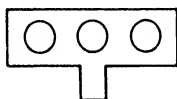




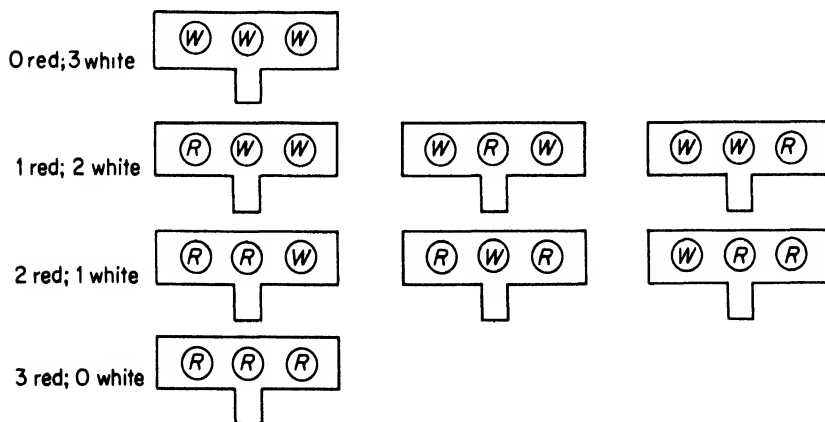
## 6.8 THE BINOMIAL DISTRIBUTION AND ITS PARAMETERS

Let an experiment be such that it can give rise to only two events, which are independent and have well defined probabilities associated with them. These two events may be called "success" and "failure." Let each repetition of the experiment be called a "trial." For example, in coin tossing a head may be a success; a tail, a failure; and a trial, a single flip of the coin. If successive trials of the experiment are independent of each other and if a trial does not change the probability of success, the trials are said to be *Bernoulli* or *binomial*. We now illustrate the binomial distribution, which many statisticians feel to be the most fundamental and important probability distribution in statistics.

**Symmetrical Distribution.** Suppose a box contains a very large number of balls, half of which are red, and the other half white. Symbolically,  $P = 0.5$  and  $Q = 1 - P = 0.5$ . Here  $P$  denotes the probability of a red ball and  $Q$  the probability of a white ball. After thorough mixing, a paddle shaped like the following is slid into the box, and a sample of three balls is scooped up.



The different samples, each of which has a probability of  $(\frac{1}{2})(\frac{1}{2})(\frac{1}{2}) = \frac{1}{8}$ , that may result are as follows:



We see that in a sample of 3 there is 1 way of getting 0 red balls; there are 3 ways of getting 1 red ball; there are 3 ways of getting 2 red balls; there is 1 way of getting 3 red balls. Or more generally, we may say

$$\binom{3}{0} = \frac{3!}{0!3!} = \frac{3 \cdot 2 \cdot 1}{(1)(3 \cdot 2 \cdot 1)} = 1$$

$$\binom{3}{1} = \frac{3!}{1!2!} = \frac{3 \cdot 2 \cdot 1}{(1)(2 \cdot 1)} = 3$$

$$\binom{3}{2} = \frac{3!}{2!1!} = \frac{3 \cdot 2 \cdot 1}{(2 \cdot 1)(1)} = 3$$

$$\binom{3}{3} = \frac{3!}{3!0!} = \frac{3 \cdot 2 \cdot 1}{(3 \cdot 2 \cdot 1)(1)} = 1$$

The values of  $\binom{n}{d}$  are often called *binomial coefficients*. Here,  $n$  represents the sample size and  $d$  the number of red balls drawn. The student who is unfamiliar with the use of binomial coefficients is referred to Appendix 17.

The symmetrical binomial distribution is obtained by multiplying each binomial coefficient by the probability  $0.5^n$  of any sample. Thus the expression for the binomial probability distribution when  $P = Q = 0.5$  is

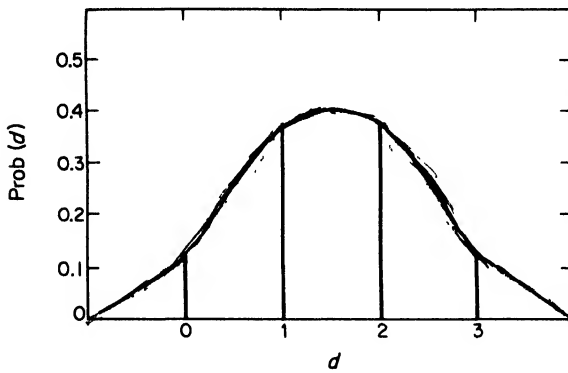
$$\text{Prob}(d) = \binom{n}{d} 0.5^n \quad (6-16)$$

See Table 6.3. The results are shown graphically in Chart 6.5. This distribution is symmetrical, for  $\text{Prob}(d=0)$  is the same as  $\text{Prob}(d=3)$  and  $\text{Prob}(d=1)$  is the same as  $\text{Prob}(d=2)$ . A binomial distribution is always symmetrical when the two classes (such as red balls and white balls) have the same number of items in the population.

**TABLE 6.3: PROBABILITY DISTRIBUTION OF NUMBER OF RED BALLS IN SAMPLES OF THREE, FROM A POPULATION WITH HALF THE BALLS RED AND THE OTHER HALF WHITE**

Number of red balls $d$	Binomial coefficient $\binom{n}{d}$	Probability of any sample, $0.5^n$	$\text{Prob}(d) = \binom{n}{d} 0.5^n$
0	1	0.125	0.125
1	3	0.125	0.375
2	3	0.125	0.375
3	1	0.125	0.125
Total	8	...	1.000

**CHART 6.5: PROBABILITY DISTRIBUTION OF NUMBER OF RED BALLS IN SAMPLES OF THREE FROM A POPULATION OF BALLS THAT IS HALF RED AND HALF WHITE.**



**Skewed Distribution.** Now let us determine the probability distribution for samples of 3 from a population that is 40 percent defective (bad) and 60 percent effective (good). For convenience, let  $P$  be the proportion of defective items in the population and  $Q$  be the proportion effective.  $P + Q = 1$ . In the present case  $P = 0.4$  and  $Q = 0.6$ . We proceed exactly as before, except that instead of multiplying  $\binom{n}{d}$  by  $0.5^n$ , we multiply it by  $P^d Q^{n-d}$ , which is the probability that a specified number of items will be bad, the others being good. Thus we have<sup>(4)</sup>

$$\text{Prob} (d) = \binom{n}{d} P^d Q^{n-d} \quad (6-17)$$

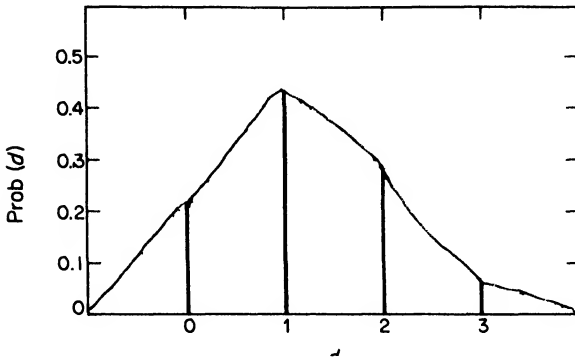
Table 6.4 shows the computations for the probability distribution. The column labeled  $p$ , the proportion defective in a sample, is not needed for

**TABLE 6.4: PROBABILITY DISTRIBUTION OF NUMBER OF DEFECTIVES IN SAMPLES OF 3 FROM A POPULATION THAT IS 40 PERCENT DEFECTIVE**

Number defective $d$	Proportion defective $p = \frac{d}{n}$	Binomial coefficient $\binom{3}{d}$	$P^d Q^{n-d}$	Prob ( $d$ ) and Prob ( $p$ )
0	0.000	1	$0.4^0 0.6^3 = 0.216$	0.216
1	0.333	3	$0.4^1 0.6^2 = 0.144$	0.432
2	0.667	3	$0.4^2 0.6^1 = 0.096$	0.288
3	1.000	1	$0.4^3 0.6^0 = 0.064$	0.064
Total	...	8	...	1.000

<sup>(4)</sup> Equation (6-16) is really a special case of Eq. (6-17), for since  $P = Q = 0.5$ , we have  $P^d Q^{n-d} = 0.5^d 0.5^{n-d} = 0.5^n$ .

**CHART 6.6: PROBABILITY DISTRIBUTION OF NUMBER OF DEFECTIVES IN SAMPLES OF THREE FROM A POPULATION THAT IS 40 PERCENT DEFECTIVE.**



computational purposes but is nevertheless of some interest. The results are shown graphically in Chart 6.6. The distribution shown in Chart 6.6 is positively skewed. If  $P$  were 0.6 and  $Q$  were 0.4, the figure would be negatively skewed. It would appear as if Chart 6.6 were viewed from the reverse side of the page.

Actually all we have done in Table 6.4 is to expand the binomial  $(Q + P)^n$ , which is accomplished by substituting the different values of  $d$  in Eq. (6-17) for the general term of the binomial. In the present case we have the binomial  $(0.6 + 0.4)^3$ , and the expression for the general term is

$$\text{Prob } (d) = \binom{3}{d} 0.4^d 0.6^{3-d}$$

Therefore,

$$\begin{aligned} (0.6 + 0.4)^3 &= \binom{3}{0} 0.4^0 0.6^3 + \binom{3}{1} 0.4^1 0.6^2 + \binom{3}{2} 0.4^2 0.6^1 + \binom{3}{3} 0.4^3 0.6^0 \\ &= 0.216 + 0.432 + 0.288 + 0.064 = 1 \end{aligned}$$

Let us now return to the question raised in the first paragraph of this chapter. Summarizing the information, we have  $P = 0.01$ ,  $Q = 0.99$ ,  $n = 50$ ,  $d = 7$ ,  $n - d = 43$ . From Eq. (6-17)

$$\begin{aligned} \text{Prob } (7) &= \binom{50}{7} (0.01)^7 (0.99)^{43} \\ &= (99,884,400)(0.0000000000000001)(0.64910) \\ &= 0.0000006 \end{aligned}$$

If we should compute the probability of obtaining 8 defectives it would be so small that we could neglect it. This fact would be true also for any number of defectives greater than 8. Since the probability of obtaining 7 or more

defectives is less than one in one million, it is unreasonable to suppose that the lot is only one percent defective. The lot should therefore be rejected.

The expected value of a binomial distribution may be calculated with considerable ease by use of

$$E(d) = nP = n(1 - Q) \quad (6-18)$$

Similarly, the variance is given by

$$\sigma_d^2 = nPQ = nP(1 - P) \quad (6-19)$$

The student should verify these "short-cut" formulas using the previously given distributions and Eqs. (6-14) and (6-15).

As we have pointed out, the skewness exhibited by a binomial distribution will depend upon the relationship between  $P$  and  $Q$ . Thus

$$\alpha_3(d) = \frac{Q - P}{\sigma_d} = \frac{Q - P}{\sqrt{nPQ}} \quad (6-20)$$

is a measure of relative skewness, and a measure of relative kurtosis is given by

$$\alpha_4(d) = \frac{1 - 6PQ}{nPQ} + 3 \quad (6-21)$$

Notice again that when

$$\begin{array}{ll} P < Q & \text{skewness is positive} \\ P = Q & \text{skewness is zero} \\ P > Q & \text{skewness is negative} \end{array}$$

Appendix 7-a gives values of selected binomial coefficients and should be of aid in calculating various binomial distributions. Extensive tables of the binomial distribution are available.<sup>(5)</sup>

## 6.9 THE POISSON DISTRIBUTION AND ITS PARAMETERS

Consider a piece of paper with blemishes on its surface. Let each of these blemishes be called a *defect*,  $c$ . The number of defects per unit usually has the Poisson distribution

$$\text{Prob}(c) = \frac{a^c e^{-a}}{c!} \quad (6-22)$$

<sup>(5)</sup> For example, Harry G. Romig, *50-100 Binomial Tables* (New York: John Wiley, 1953). Also, U.S. Dept. of Commerce, National Bureau of Standards, Applied Mathematics Series 6, *Tables of the Binomial Probability Distribution*, 1950. Or Harvard University (Computation Laboratory), *Tables of the Cumulative Binomial Probability Distribution*, (Cambridge, Mass.: Harvard University Press, 1965).

TABLE 6.5: CALCULATION OF A POISSON DISTRIBUTION FOR  $a = 0.5$ 

$c$	$a^c$	$e^{-a}$	$c!$	$Prob(c) = a^c e^{-a} / c!$	<i>Cumulative Prob(c)</i>
0	1.0	0.6065*	1	0.6065	0.6065
1	0.5	0.6065	1	0.3033	0.9098
2	0.25	0.6065	2	0.0758	0.9856
3	0.125	0.6065	6	0.0126	0.9982
4	0.0625	0.6065	24	0.0016	0.9998
5	0.03125	0.6065	120	0.0002	1.0000

\*See Appendix 7-b.

where  $c$  is the number of defects per sample,  $a$  is the expected number of defects per sample, and  $e \doteq 2.71828$ , which is the base of the natural logarithm system.

The Poisson distribution is the limit of the binomial as  $n$  approaches infinity with  $nP$  constant. To explain this, we may think of the defects as being distributed over the area of a surface, any unit being defective if it contains one or more defects. If the surface area is divided into very small units, so that no unit of area has more than one defect, the distinction between defect  $c$  and defective  $d$  disappears. As the number of units  $n$  increases,  $nP$  remaining constant,  $P$  must get smaller and smaller, and the binomial distribution gradually approaches the Poisson.

Table 6.5 illustrates the calculation of a Poisson distribution for  $a = 0.5$  from  $c = 0$  through 5.

The expected value and variance of a Poisson distribution are both given by the single formula

$$E(c) = \sigma_c^2 = a \quad (6-23)$$

The Poisson distribution is always positively skewed, but the skewness diminishes as  $a$  increases

$$\alpha_3(c) = \frac{1}{\sigma_c} = \frac{1}{\sqrt{a}} \quad (6-24)$$

and kurtosis is measured by

$$\alpha_4(c) = \frac{1}{a} + 3 \quad (6-25)$$

Since the Poisson distribution is considerably simpler to compute for large values of  $n$ , it is often used as an approximation to the binomial when  $n$  is very large and  $P$  is very small. Then

$$Prob(d) \doteq \frac{a^d e^{-a}}{d!} \quad (6-26)$$

where  $a = nP$  and  $d$  is the number of defective items in a sample. A workable rule of thumb for substituting the Poisson distribution for the binomial distribution is as follows: when  $n > 10$  and  $P < 0.1$ , the Poisson distribution

may be used to approximate the binomial distribution with negligible error. The student may verify this rule through experimentation.

Appendix 7-b gives selected values of  $e^{-a}$  and should be of aid in calculating various Poisson distributions. Extensive tables of the Poisson distribution are available.<sup>(6)</sup>

The Poisson distribution is useful in ways other than as an approximation to the binomial; for example, the analysis of queues, or arrival and waiting line patterns at bridges and airports, tool distribution points in factories, and so on.

## 6.10 OTHER TYPES OF DISCRETE PROBABILITY DISTRIBUTIONS

Theoretically, there are infinitely many discrete probability distributions. Of these, only a small number have been fully explored by statisticians, and the binomial and Poisson distributions are the two most often encountered in both practical and theoretical statistical investigation. Some other types of discrete probability distributions are

1. The hypergeometric.
2. The negative binomial.
3. The geometric.
4. The multinomial.

The hypergeometric distribution is used instead of the binomial in cases where the sample size is large relative to the population. It is discussed in Sec. 12.3.

## 6.11 BAYES' THEOREM

Suppose that  $A$  and  $B$  are mutually exclusive events that exhaust the sample space of an experiment associated with them, for example, effective and defective items. Now let  $d$  be the actual outcome of an experiment. We wish to find  $\text{Prob}(A|d)$  and/or  $\text{Prob}(B|d)$ . For example, suppose that there are two machine types called  $A$  and  $B$ . Twenty-five percent of the machines in a factory are of type  $A$  and 75 percent of type  $B$ . Machine type  $A$  is known to produce items that, in the long run, are 5 percent defective. Machine type  $B$  produces items that, in the long run, are 10 percent defective. The output from these machines in a given day is 1000 units, and an item drawn at random from this population is found to be defective. What is the

<sup>(6)</sup> Among the best known is E. C. Molina, *Poisson's Exponential Binomial Limit*, (Princeton, N.J.: D. Van Nostrand, 1942).

probability that it was produced by machine type  $A$ ? Or alternatively, what is the probability that it was produced by machine type  $B$ ? Thus, if  $d$  represents a defective item, we seek

$$\text{Prob}(A | d) \text{ and/or } \text{Prob}(B | d)$$

where  $A$  and  $B$  represent the two machine types.

Using Eq. (6-7), we may write for our problem

$$\text{Prob}(A | d) = \frac{\text{Prob}(A \cap d)}{\text{Prob}(d)}$$

Also, from Eq. (6-8)

$$\text{Prob}(A \cap d) = \text{Prob}(A) \cdot \text{Prob}(d | A)$$

But  $\text{Prob}(d)$  is the sum of the joint probabilities

$$\begin{aligned} \text{Prob}(d) &= \text{Prob}(A \cap d) + \text{Prob}(B \cap d) \\ &= \text{Prob}(d | A) \cdot \text{Prob}(A) + \text{Prob}(d | B) \cdot \text{Prob}(B) \end{aligned}$$

Therefore

$$\text{Prob}(A | d) = \frac{\text{Prob}(d | A) \cdot \text{Prob}(A)}{\text{Prob}(d | A) \cdot \text{Prob}(A) + \text{Prob}(d | B) \cdot \text{Prob}(B)} \quad (6-27)$$

Equation (6-27) is often called Bayes' theorem in honor of the 18th-century clergyman and mathematician, T. Bayes.<sup>(7)</sup>

Let us illustrate the solution to our problem using Bayes' theorem. We define the following probabilities from the discussion above:

$$\begin{aligned} \text{Prob}(A) &= 0.25 & \text{Prob}(B) &= 0.75 \\ \text{Prob}(d | A) &= 0.05 & \text{Prob}(d | B) &= 0.10 \end{aligned}$$

$$\text{Then} \quad \text{Prob}(A | d) = \frac{(0.05)(0.25)}{(0.05)(0.25) + (0.10)(0.75)} = \frac{1}{7}$$

$$\text{and} \quad \text{Prob}(B | d) = \frac{(0.10)(0.75)}{(0.10)(0.75) + (0.05)(0.25)} = \frac{6}{7}$$

$$\text{Notice that} \quad \text{Prob}(B | d) + \text{Prob}(A | d) = 1$$

---

<sup>(7)</sup> Bayes' theorem may be extended in the following manner. Given the events  $A, B, C, \dots$ ,

$$\text{Prob}(A | d) = \frac{\text{Prob}(d | A) \cdot \text{Prob}(A)}{\text{Prob}(d | A) \cdot \text{Prob}(A) + \text{Prob}(d | B) \cdot \text{Prob}(B) + \text{Prob}(d | C) \cdot \text{Prob}(C) + \dots}$$



## PROBLEMS

1. If  $n$  fair coins are tossed together the distribution of heads  $d$  is given by

$$\text{Prob}(d) = \frac{\binom{n}{d}}{2^n}, \quad d = 0, 1, 2, \dots, n$$

Find the probability of receiving 3 heads or fewer in a single toss of 4 fair coins. Find the expected value and variance of this distribution.

2. A large batch of nuts and bolts are on an assembly table. Fifty percent of the nuts are defective, 20 percent of the bolts are defective. Of the effective nuts, 10 percent will fit the effective bolts. Find the probability that an effective nut and an effective bolt which will fit each other will be drawn at random.

3. The probability that an operator of a given machine will be injured by the machine in a given time period is 0.001. If 1000 operators are at work during this time period, what is the probability that 2 or fewer will suffer an injury? (We recommend using the Poisson distribution to approximate the binomial distribution in this problem. Why?)

4. A lottery offers 5 prizes with values and probabilities given below.

Prize	Value	Prob (winning)
First	\$100	0.001
Second	50	0.010
Third	25	0.020
Fourth	10	0.030
Fifth	5	0.040

The column headed Prob (winning) gives the probability that a single ticket will be drawn at random from a very large number of tickets and awarded a prize. Assuming that you should not pay more than the expected value of your winning for a ticket, would you purchase a ticket costing \$1.00?

5. Two types of coins are in an urn. Thirty percent are fair, with Prob (heads) = 0.5, and 70 percent are unfair, with Prob (heads) = 0.8. A coin is drawn at random from this urn and tossed. Heads appears. What is the probability that the coin is fair? Unfair?

6. In the questions for Chapter 7 you will encounter the name "Chebyshev." One of the authors is puzzled about the English spelling of this name. Upon looking in various references he found the following:

- a. *CH* is sometimes used in place of *SH*.
- b. *T* sometimes precedes the initial *CH*.
- c. *FF* is sometimes used in place of the final *V*.

If all variations of spelling are equally probable, what is the probability that a given reference will spell the name Tchebycheff?

7. Write Eq. (6-11) for five events.

## The Normal Probability Distribution

In the last chapter we discussed some properties of discrete probability distributions and indicated that these distributions were associated with experiments which could give rise only to outcomes that were discrete in nature. In Chapter 2 we distinguished between discrete and continuous variables, and in this chapter we continue along these lines to treat experiments which may generate a continuous set of outcomes and which have associated with them continuous probability distributions. Of the infinite number of possible continuous probability distributions we will for the present treat only one, namely, the “normal” probability distribution. The word “normal” should not be taken to mean “usual” or “typical.” Normal is a generic name applied to a well defined continuous probability distribution, which is also called Gaussian or, more rarely, Laplacean (or some variation on the names of the mathematicians Gauss and Laplace). Other important continuous probability distributions such as the “Student’s”  $t$ , the chi-square, and the variance ratio, or  $F$  distribution, will be treated in later chapters.

In previous chapters curves were shown which illustrated some of the forms that a frequency distribution may assume. These curves were based upon data of a few score or a few hundred cases; each was a sample from a much larger, possibly infinite, population. Being a sample, a given curve would not necessarily have exactly the same shape as the curve for the population, but if the sample is properly selected, the curve for the sample will tend to be of the same general shape as the curve for the population. A curve based

upon sample data will show certain irregularities, but we may fit a curve to the data obtained and thus smooth out those irregularities presumably attributable to sampling. Such a fitted curve is a generalization, believed to represent the actual situation underlying the sample.

Many types of curves may be fitted to frequency distributions. For example, when dealing with data representing a continuous variable, we may fit a symmetrical or an asymmetrical curve of the Pearsonian group or of the Gram-Charlier series,<sup>(1)</sup> whereas for data representing a discrete variable, we may use a hypergeometric, a binomial, or a Poisson distribution. For the purposes of this text we will discuss only one of the symmetrical curves of the Pearsonian group, the "normal" curve. This is also a "generating function" for the Gram-Charlier series.

We may be interested in fitting a curve to a set of data in order to generalize concerning the fundamental shape of the distribution. Such a purpose is served when a normal curve is used to describe errors made in repeated measurements. This use also permits us to estimate the proportion of measurements that will fall within a certain range above, below, or between selected values.

If we are studying the life expectancy of physical property, such as telephone poles, it may be important to know what proportion of the poles will need to be replaced during each year after installation. An example of this sort will be considered in an appendix to this chapter. A curve may also be fitted to one set of data in order to generalize concerning an associated variable. Thus a curve may be fitted to the circumferences of boys' heads, enabling a reasonable conclusion to be drawn as to the number of caps of each size that should be made for such a group of boys. Finally, one of the most important uses of the normal distribution is in the making of inferences concerning the arithmetic mean. This use will be discussed in a later chapter.

## 7.1 NORMAL CURVE AS LIMITING FORM OF OTHER DISTRIBUTIONS

The normal curve represents a distribution of values that may occur, under certain conditions, when chance is given full play. In every case the necessary conditions include the existence of a large number of causes, each operating independently in a random manner.

**Symmetrical Binomial.** The operation of chance may be illustrated by coin tossing. For coin-tossing experiments, the coin should be

---

<sup>(1)</sup> For a discussion of the Pearsonian and Gram-Charlier systems, as well as illustration of fitting, see Dudley J. Cowden, *Statistical Methods in Quality Control* (Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1957), Chap. 20.

so constructed that it is incapable of standing on edge, and is perfectly balanced, so that a head or a tail is equally probable. Thus, if one coin is tossed, the probability of throwing a tail or a head is indicated by

$$\frac{1}{2}T + \frac{1}{2}H = 0.5T + 0.5H$$

The probability distribution is rectangular, the probability of 0 heads or 1 head each being 0.5.

If two coins are tossed the probabilities are represented by

$$(\frac{1}{2}T + \frac{1}{2}H)^2 = \frac{1}{4}T^2 + \frac{2}{4}TH + \frac{1}{4}H^2 = 0.25T^2 + 0.50TH + 0.25H^2$$

The exponents of  $T$  and  $H$  indicate the number of tails or heads. Thus, the probability is 0.25 of obtaining two tails, 0.50 of obtaining a tail and a head, and 0.25 of obtaining two heads. Or we may say that the probability of 0 heads is 0.25, of 1 head is 0.50, of 2 heads is 0.25. The probability distribution is thus triangular in shape. If 2 coins are thrown 1000 times, our expectation would be 0 heads 250 times, 1 head 500 times, and 2 heads 250 times.

If four coins are tossed, the probabilities are

$$\begin{aligned} (\frac{1}{2}T + \frac{1}{2}H)^4 &= \frac{1}{16}T^4 + \frac{4}{16}T^3H + \frac{6}{16}T^2H^2 + \frac{4}{16}TH^3 + \frac{1}{16}H^4 \\ &= 0.0625T^4 + 0.25T^3H + 0.375T^2H^2 + 0.25TH^3 + 0.0625H^4 \end{aligned}$$

These binomial probabilities are of course obtained by use of the formula

$\text{Prob}(d) = \binom{n}{d}P^dQ^{n-d}$ . In these expressions  $d$  refers to the number of heads.

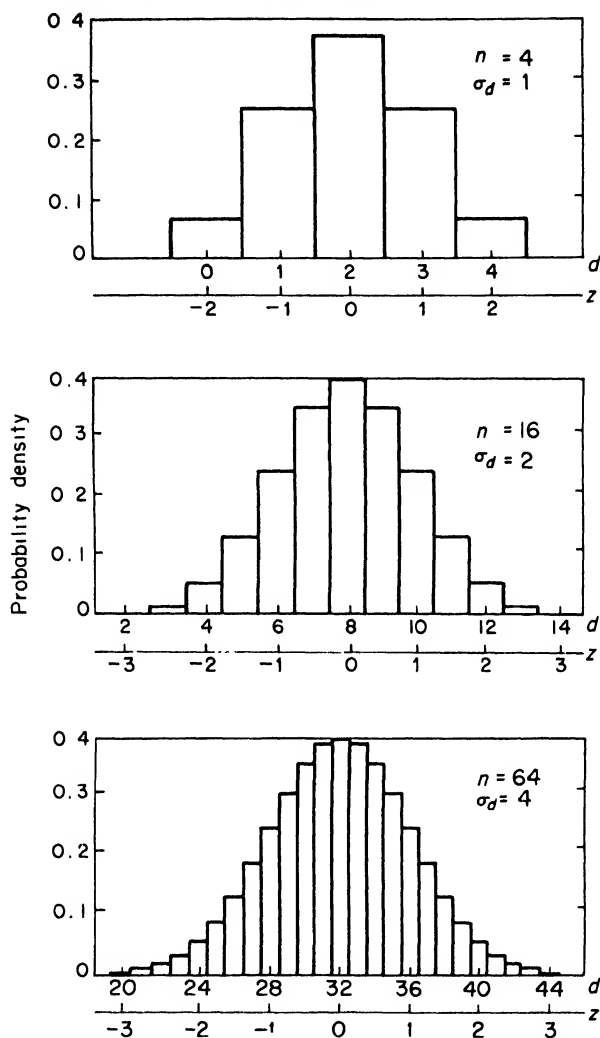
The results are recorded in the table below.

#### BINOMIAL PROBABILITY DISTRIBUTION

$$n = 4, \quad P = 0.5$$

<i>Number of heads d</i>	<i>Prob (d)</i>
0	0.0625
1	0.2500
2	0.3750
3	0.2500
4	0.0625
Total	1.0000

As  $n$  approaches infinity, the binomial distribution approaches normal form. Chart 7.1 illustrates the gradual approach of the binomial  $(0.5T + 0.5H)^n$  to the normal form as  $n$  is in turn 4, 16, and 64. In each case the horizontal scale is so selected that the different histograms will exhibit the same amount of dispersion in terms of  $z$ . Technically, each distribution has

CHART 7.1: BINOMIAL PROBABILITY DISTRIBUTION,  $P = 0.5$ .

been standardized, and the horizontal scales are given both in terms of  $d$  and  $z$ , where

$$z = \frac{d - E(d)}{\sigma_d} = \frac{d - nP}{\sqrt{nPQ}}$$

In the *upper* section of Chart 7.1 we see that  $z$  may take on only integral values. Each column in this histogram is centered on one of these  $z$  values. The columns are each one unit wide both in terms of  $d$  and  $z$ . The probability associated with a given value of  $d$  or  $z$  is simply the height of the given column.

The distribution depicted in the *center* section of Chart 7.1 has a standard deviation twice as large as the distribution shown in the upper section, and twice as many columns are allocated to a standard deviation. In order to keep the total area in the two sections the same, we have made the columns twice as high as that obtained by the binomial expansion. Thus,  $\text{Prob}(d=8) = 0.1964$ , but the height of the central bar, the *probability density*, is  $2(0.1964) = 0.3928$ . In terms of the  $z$  values, each bar in the center section is half as wide as a bar in the upper section. In general, the width of the bars in terms of  $z$  units is  $1/\sigma_d$ . Since the probability of a given  $z$  value is the area of the associated column,

$$\begin{aligned}\text{Prob}(z) &= \text{height of column} \cdot \text{width of column} \\ &= \text{prob. density}(z) \cdot 1/\sigma_d\end{aligned}$$

or,<sup>(2)</sup> when  $z = 0$ ,

$$\text{Prob}(z=0) = (0.3928)\frac{1}{2} = 0.1964$$

In the *lower* section of Chart 7.1 the columns are one-fourth of a  $z$  unit wide and, therefore, the columns are four times as tall as the binomial probabilities.

The three column diagrams exhibit considerable similarity in appearance. The chief difference is that as  $n$  becomes larger, the bars become narrower in terms of  $z$  units and more numerous. If  $n$  were to be continually increased, the bars would become narrower and narrower in terms of  $z$  units until the steps would finally disappear, and we would have a continuous curve that is normal in form. See Chart 7.2.

It can be shown that as  $n$  approaches infinity, with  $P$  constant, the standardized binomial distribution approaches the standardized normal distribution. The formula for the normal distribution is<sup>(3)</sup>

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(X-\mu)^2/2\sigma^2} \quad (7-1)$$

The symbol  $f(X)$  is read "function of  $X$ ," or simply " $f$  of  $X$ ." Equation (7-1) describes a normal probability distribution with mean  $\mu$  and variance  $\sigma^2$  (standard deviation  $\sigma$ ).

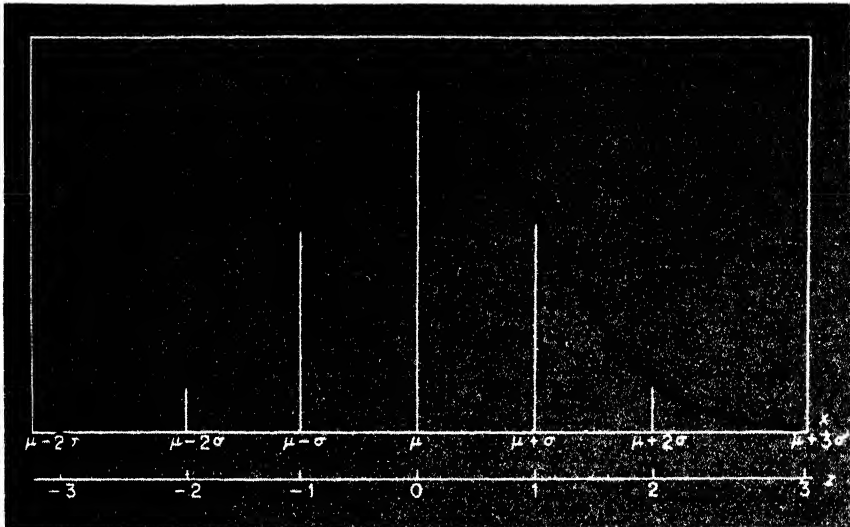
Equation (7-1) can be simplified if we consider the normal distribution as a function of  $z$ , the standard measure, where  $z = (X - \mu)/\sigma$ .

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \quad (7-2)$$

<sup>(2)</sup> In general,  $\text{Prob. density}(d) = \sigma_d \cdot \text{Prob}(d)$

<sup>(3)</sup>  $\pi$  is the ratio of the circumference to the diameter of a circle, whereas  $e$  is the base of the natural, or less aptly Napierian, system of logarithms.

$$e = \lim_{X \rightarrow \infty} \left(1 + \frac{1}{X}\right)^X \doteq 2.71828; \quad \pi \doteq 3.14159$$

CHART 7.2: NORMAL CURVE EXPRESSED AS A FUNCTION OF BOTH  $X$  AND  $z$ .

In this form, of course, the distribution has a mean of zero and a variance of one.

Chart 7.2 shows a normal curve expressed both as a function of  $X$  and as a function of  $z$ . Notice that the maximum ordinate of this curve is located at  $X = \mu$  when the curve is expressed as in Eq. (7-1). When the curve is expressed in standard form as in Eq. (7-2), the maximum ordinate is at  $z = 0$  and is

$$f(0) = \frac{1}{\sqrt{2\pi}} \quad (7-3)$$

One important mathematical property of any normal curve is that it *never* touches the horizontal axis no matter how large or small  $X$  or  $z$  become. This property, of course, cannot be shown diagrammatically. The normal curve is, therefore, said to be *asymptotic* with regard to the horizontal axis.

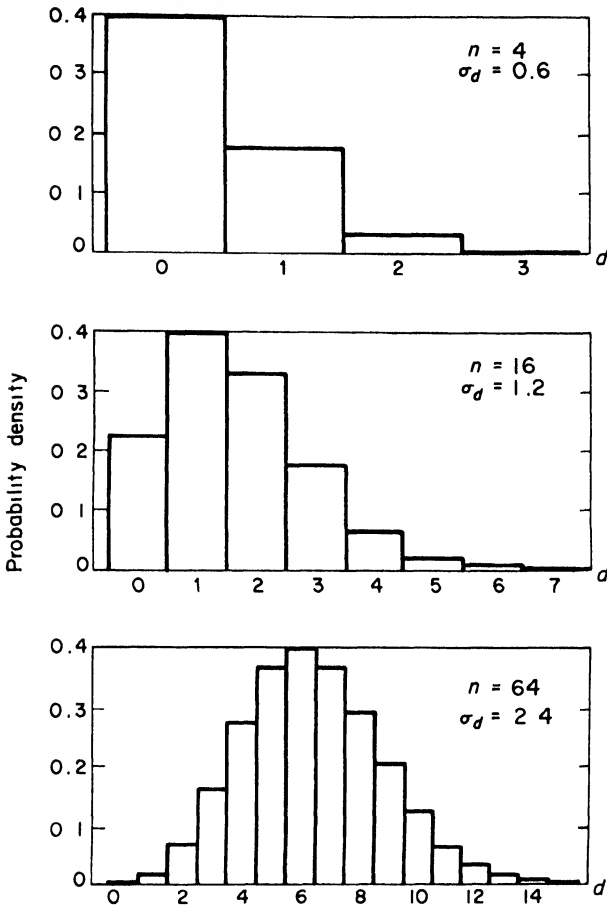
**Skewed Binomial.** If we select samples from an infinite population that is (say) 10 percent defective, we have the binomial

$$(0.9g + 0.1d)^n$$

where  $d$  is the number of defectives in a sample of size  $n$ , and  $g$  is the number of good items in the sample;  $g + d = n$ . The probability of obtaining any specified number of defectives is

$$\text{Prob}(d) = \binom{n}{d} 0.1^d 0.9^{n-d}$$

Histograms of the probability distributions when  $n = 4$ , 16, and 64 are

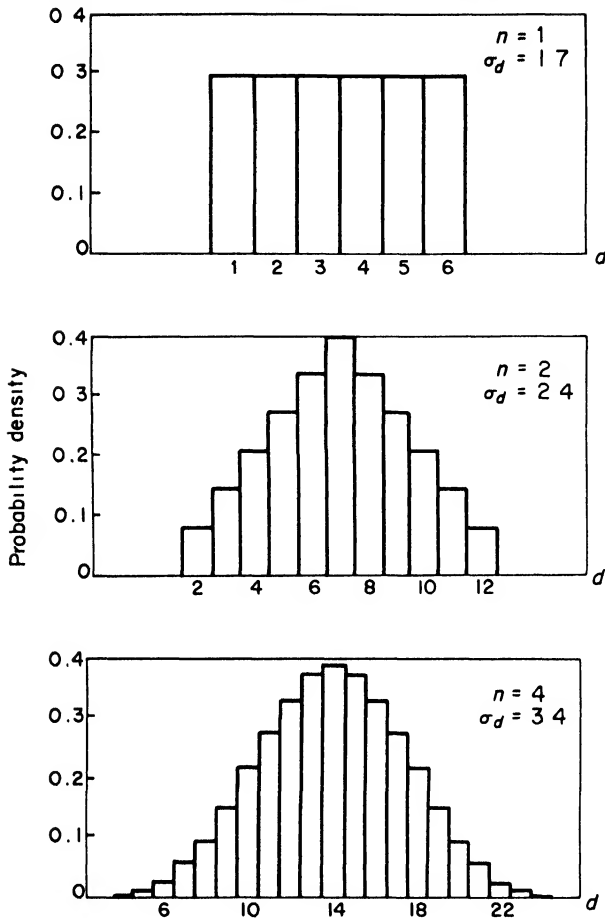
CHART 7.3: BINOMIAL PROBABILITY DISTRIBUTION,  $P = 0.1$ .

shown in Chart 7.3. This chart is constructed similarly to Chart 7.1. All generalizations made about Chart 7.1 can be applied to Chart 7.3. Also, the histograms are becoming less skewed. As  $n$  increases, the skewed binomial distribution approaches normal form, though *not so rapidly* as the symmetrical binomial.

**Tossing Dice.** If the probabilities associated with throwing a six-sided die are graphed, they will form a rectangular probability distribution. When two dice are thrown at a time, the probability distribution is triangular. When four dice are thrown at a time, however, the distribution of the total number of spots is almost normal. The rapid approach to the normal form is shown vividly by Chart 7.4.



CHART 7.4: DICE THROWING PROBABILITY DISTRIBUTION.



**The Central Limit Theorem.** We have shown graphically how three kinds of probability distributions approach the normal form. Actually we can make a much broader generalization, known as the central limit theorem. This theorem, which is perhaps the most important one in statistics, is sometimes stated as follows:

If a population has a finite variance  $\sigma^2$  and mean  $\mu$ , then the standardized distribution of the sample mean approaches the normal distribution with mean of zero and variance of one as the sample size increases.

When the sample size is as small as 4 or 5, means of random samples from populations of a continuous variable that are likely to be encountered by the

business statistician will be distributed almost normally. We can thus use normal probabilities in making statements concerning the arithmetic mean.

## 7.2 PROBABILITY AND THE NORMAL CURVE

In future chapters we will make a great deal of use of the mathematical properties of the normal distribution. In particular, we will need to know the areas under the normal curve associated with various regions of the curve. Since the normal distribution is a probability distribution, we know that the area under the entire curve is 1. Since the curve is symmetrical, we reason that the area on either side of  $\mu$  is 0.5, or 50 percent of the total area.

The next logical question might be: Given a normal distribution with mean  $\mu$  and variance  $\sigma^2$ , what percentage of the total area lies to the right or the left of a given  $X$  value? Alternatively we might ask: What is the probability of obtaining a value of  $X$  that is as large or larger than one specified? Since there are infinitely many normal curves, each depending upon a particular combination of  $\mu$  and  $\sigma^2$ , the answer would vary from normal curve to normal curve. The problem would be much less difficult if we could cause all normal distributions to have the same mean and variance. This is exactly what we do when we standardize distributions; i.e., we cause them to have a mean of zero and a variance of one. Thus, if we transform the given  $X$  value into a  $z$  value, where  $z = (X - \mu)/\sigma$ , and call  $Q(z)$  the probability of obtaining a value of  $z$  that is equal to or larger than the one specified, tables are already calculated to allow us to find  $Q(z)$ , given  $z$ . Such a table is given in Appendix 1. We now illustrate the use of this and two other related appendices.

Assume that a variable  $X$  is distributed normally with  $\mu = 5$  and  $\sigma = 2$  and that we draw at random a given  $X$  value.

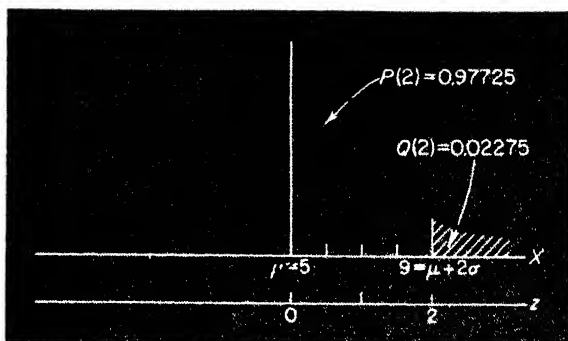
1. What is the probability of obtaining a value of  $X$  as large or larger than 9? First, we standardize the specified  $X$  value.

$$z = \frac{X - \mu}{\sigma} = \frac{9 - 5}{2} = 2$$

Entering Appendix 1, we find that  $Q(z)$ , the probability of obtaining a value of  $z$  as large or larger than specified, is

$$Q(z) = Q(2) = 0.02275$$

Alternatively, we may say that 2.275 percent of the area in the distribution is to the right of  $z = 2$ .  $Q(2)$  is shown in the chart below. The horizontal axis shows both the  $X$  variate and the corresponding  $z$  variate. The distribution is not drawn to scale.



2. What is the probability of obtaining a value of  $X$  smaller than 9? If we define  $P(z)$  as the probability of obtaining a value of  $z$  smaller than the one specified,

$$P(z) = 1 - Q(z)$$

Therefore,  $P(2) = 1 - Q(2) = 1 - 0.02275 = 0.97725$

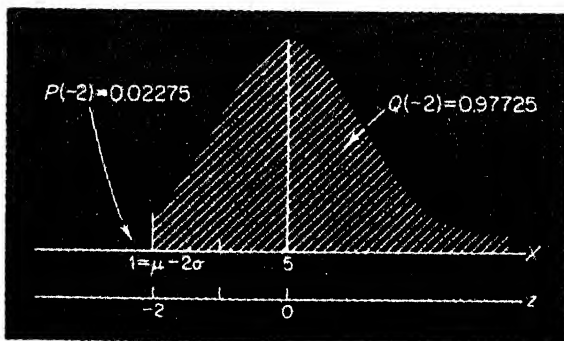
Alternatively, we may say that 97.725 percent of the area in the distribution is to the left of  $z = 2$ .  $P(2)$  is shown in the chart above.

3. What is the probability of obtaining a value of  $X$  as large or larger than 1? Here  $z = (1 - 5)/2 = -2$ . Only positive values of  $z$  are given in Appendix 1, since it is apparent that, because of the symmetry of the normal curve about  $\mu$ ,

$$Q(-z) = 1 - Q(z) = P(z)$$

Thus  $Q(-2) = 1 - Q(2) = 1 - 0.02275 = 0.97725$

$Q(-2)$  is shown below.



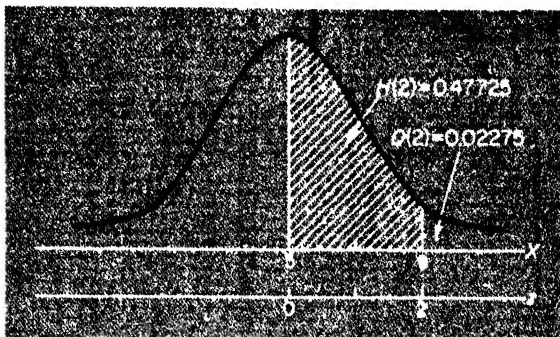
4. What is the probability of obtaining a value of  $X$  between the mean and a specified  $X$  value? For example, if  $X = 9$  and we call  $H(z)$  the area between

$\mu$  and the specified value,  $X = 9$ ,

$$H(z) = 0.5 - Q(z)$$

or 
$$H(2) = 0.5 - Q(2) = 0.5 - 0.02275 = 0.47725$$

$H(2)$  is shown below.

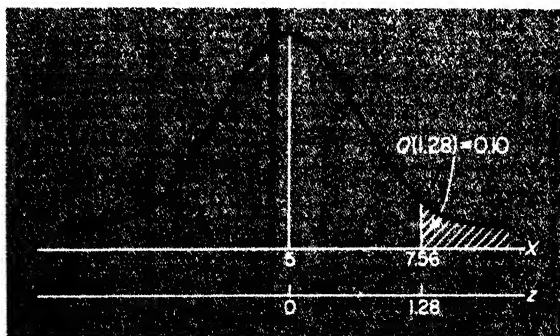


It is more convenient to use Appendix 2, where  $H(z)$  is given directly. It is also apparent, because of symmetry, that

$$H(z) = H(-z)$$

5. What is the value of  $X$  such that the probability of obtaining a value of  $X$  as large or larger than this value is 0.10? Using Appendix 1, locate the desired value of  $Q(z)$  in the body of the table. Then, in the stub, we locate  $z_Q = z_{0.10} = 1.28$ . The subscript on  $z$  is used to denote the value of  $z$  corresponding to the desired probability level. It is now a simple matter to solve for  $X$  when  $z_{0.10}$ ,  $\mu$ , and  $\sigma$  are known, since

$$z = \frac{X - \mu}{\sigma}; \text{ then } X = \mu + z\sigma$$



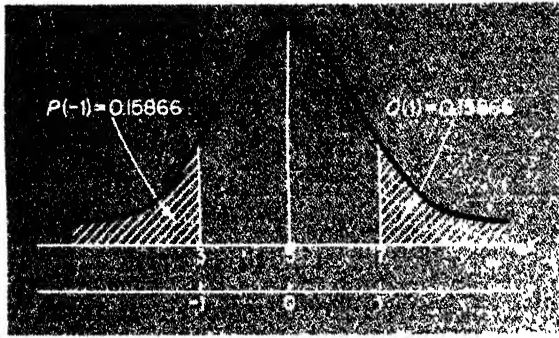
or

$$X = 5 + 1.28(2) = 7.56$$

It is more convenient to use Appendix 3, where values of  $z_Q$  are given directly for specified values of  $Q(z)$ . Hence, in Appendix 3 if  $Q(z) = 0.10$ , then  $z_Q = 1.28$ .

6. What is the probability of obtaining a value of  $z$  that is larger in absolute value (sign disregarded) than the one specified? If we specify that  $|z| = 1$ , then the area to the right of  $z = 1$  is

$$Q(1) = 0.15866$$



and to the left of  $z = -1$  is

$$P(-1) = 1 - Q(-1) = Q(1) = 0.15866$$

The total area is

$$Q(z) + P(-z) = 2(0.15866) = 0.31732$$

## PROBLEMS

1. Some sources of statistical tables give only  $Q(z)$ , whereas others give only  $H(z)$ . In view of this, work each of the following exercises, using (a) only  $Q(z)$ ; (b) only  $H(z)$ ; (c) a combination of  $Q(z)$ ,  $H(z)$ , and  $z_Q$  as the problem dictates.

Given that a random variable  $X$  is distributed normally with a mean of 10 and a variance of 4,

a. Find the probability that  $X$  assumes a value:

- i. 12 or greater.
- ii. Between 10 and 11.
- iii. Less than 9.
- iv. 8 or greater.

b. Find a value of  $X$  such that:

- i. 20 percent of the  $X$  values are as great or greater than this value of  $X$ .
- ii. 70 percent of the  $X$  values are smaller than this value of  $X$ .

c. Find  $X$ , given that:

- i.  $z = -2.0$ .
- ii.  $z = 0.0$ .
- iii.  $z = 1.0$ .

2. Assume that the results of a statistics examination are distributed normally with mean of 70 points and variance of 64. The instructor wishes to give the following grades:

- a. 10 percent F's.
- b. 15 percent D's.
- c. 60 percent C's.
- d. 10 percent B's.
- e. 5 percent A's.

Find the minimum grade necessary to make a D, C, B, and A, respectively.

3. Review the discussion of the central limit theorem (Sec. 7.1). Discuss the meaning of the terms:

- a. "Finite variance."
- b. "Distribution of the sample mean."
- c. "Approaches the normal distribution."

4. Why do we often express the normal distribution as a function of  $z$  rather than as a function of  $X$ ?

5. A very famous result in mathematical statistics is called Chebyshev's inequality. It states that for any well defined probability distribution, the proportion of values found beyond  $\pm K$  standard deviations from the mean of the distribution cannot be greater than  $1/K^2$ . Thus, for any distribution

$$\text{Prob } (|X - \mu| > K\sigma) < \frac{1}{K^2}$$

Verify this result for the normal distribution, using  $K = \pm 1, \pm 2$ , and  $\pm 3$ .

---

## APPENDIX : Using the Normal Curve to Describe an Observed Frequency Distribution

The equation

$$f(X) = \frac{1}{\sigma \sqrt{2\pi}} e^{-(X-\mu)^2/2\sigma^2}$$

describes a normal curve with mean  $\mu$  and variance  $\sigma^2$  in terms of probability density per class with interval of  $1\sigma$ . If we have  $n$  sample observations, we wish the area under the curve to be  $n$ , and since the frequency densities of any distribution vary directly with the class interval  $c$ , we must substitute  $nc$  for 1 in the above expression when fitting a normal curve to sample data. Furthermore, we must substitute  $\bar{X}$  for  $\mu$  and  $SD$  for  $\sigma$ , in the above expression, since  $\bar{X}$  and  $SD$  are joint maximum likelihood estimators of  $\mu$  and  $\sigma$ . Thus, we have

$$f(X) = \frac{nc}{SD} \frac{1}{\sqrt{2\pi}} e^{-(X-\bar{X})^2/2(SD)^2}$$

$\hat{f}(X)$  is used instead of  $f(X)$  in order to indicate that we are referring to a curve fitted to a sample, rather than the curve of a population. The maximum ordinate,  $\hat{f}(\bar{X})$ , is found when  $X = \bar{X}$  and is

$$\hat{f}(\bar{X}) = \frac{nc}{SD} \frac{1}{\sqrt{2\pi}}$$

since  $e^0 = 1$ . By substitution we may write

$$\hat{f}(X) = \hat{f}(\bar{X}) e^{-(X-\bar{X})^2/2(SD)^2}$$

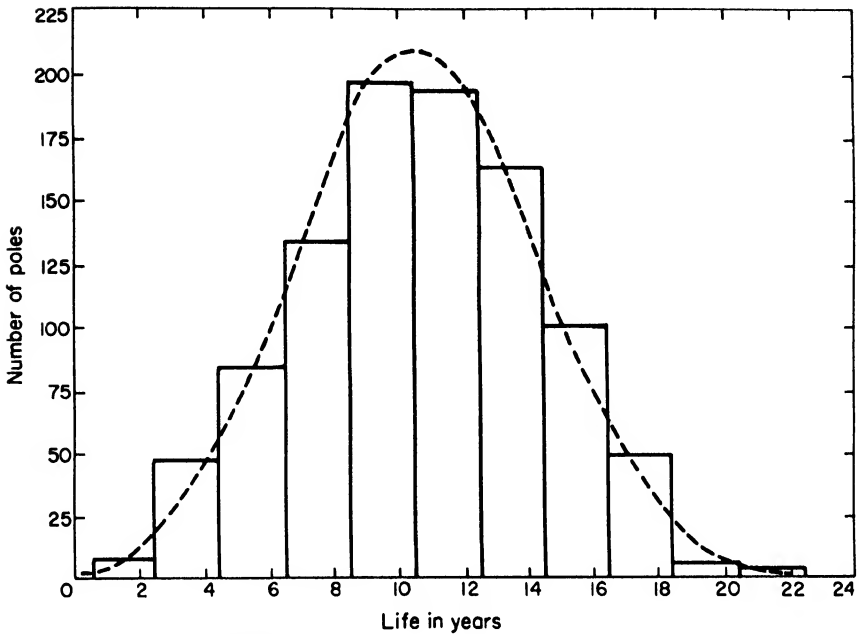
It is, of course, true that the items in a sample cannot be distributed normally, since  $n$  is a finite number and an actual sample cannot extend to  $\pm\infty$ . But  $\hat{f}(X)$  refers to the expected frequency densities of a sample of size  $n$  taken from a normal population with the same mean and standard deviation as the sample.

Table A7.1 shows the life experience of 1000 wooden telephone poles. If a

**TABLE A7.1: LIFE EXPERIENCE OF WOODEN TELEPHONE POLES (YEARS)**

<i>Class limits</i>	<i>Mid-value</i>	<i>Number replaced</i>
0.5-2.5	1.5	11
2.5-4.5	3.5	47
4.5-6.5	5.5	87
6.5-8.5	7.5	134
8.5-10.5	9.5	200
10.5-12.5	11.5	198
12.5-14.5	13.5	164
14.5-16.5	15.5	102
16.5-18.5	17.5	48
18.5-20.5	19.5	6
20.5-22.5	21.5	3
Total	...	1000

Source: Adopted from Roble Winfrey and Edwin B. Kurtz *Life Characteristics of Physical Property*, Bulletin 103, Iowa Engineering Experiment Station, p. 57, property group 24-5.

**CHART A7.1: HISTOGRAM OF LIFE EXPERIENCE OF 1,000 WOODEN TELEPHONE POLES AND PLOTTED FREQUENCY CURVE.**

Source: Table A7.1.

suitable curve can be fitted to the data, it will be possible to state the expected proportion or number of poles to be replaced each year. The histogram of Chart A7.1 indicates that a normal curve might describe the data well. Therefore, we shall use the normal curve to estimate the number of poles that will wear out during specified intervals of time.

The computation of theoretical frequencies in each class interval consists essentially of integrating the normal curve. The process of integration is avoided, however, by making use of a table of normal curve probabilities, such as Appendix 1. The procedure, as illustrated in Table A7.2 and indicated symbolically at the top of the different columns, is as follows:

1. Record the class limits. Each of the  $X$  values except  $-\infty$  and  $+\infty$  is both an upper limit of one class and a lower limit of the next.
2. Subtract the mean (which is 10.658) from each class limit, obtaining deviations from the mean.
3. Divide each of these deviations by the standard deviation (which is 3.765), thus converting the class limits into standard measures.
4. Ascertain from Appendix 1 the value of  $Q(z)$ .
5. To estimate the probability of a pole's lasting at least  $X$  years, subtract each  $Q(z)$  value from the one immediately above it and record the difference in the space between the class limits. Thus, the probability is 0.99653 that a



pole will last more than 0.5 years and 0.98500 that a pole will last more than 2.5 years, so the probability is  $0.99653 - 0.98500 = 0.01153$  that a pole will last between 0.5 year and 2.5 years. In the column for  $\hat{f}/n$ , the subscript 1 is used to refer to the lower limit of any class and the subscript 2 is used to refer to the upper limit of that class.

6. The total of the  $\hat{f}/n$  column is 100 percent. Since there were 1000 poles in the original distribution, the probabilities of the  $\hat{f}/n$  column are multiplied by  $n$  (which is 1000) to give the expected frequencies of the last column.

**TABLE A7.2: COMPUTATION OF EXPECTED FREQUENCIES IN EACH CLASS FOR LIFE OF WOODEN POLES ( $\bar{X} = 10.658$  YEARS;  $SD = 3.765$  YEARS;  $n = 1000$ )**

<i>Class limits (years)</i>	<i>Deviation from mean <math>x</math></i>	<i>Standard measure <math>z</math></i>	<i>Probability of lasting at least <math>X</math> years*</i>	<i>Probability of wearing out between stated time limits <math>\hat{f}/n</math></i>	<i>Expected number wearing out between stated time limits <math>\hat{f}</math></i>
$X$	$X - \bar{X}$	$\frac{x}{SD}$	$Q(z)$	$Q(z_1) - Q(z_2)$	$n \frac{(\hat{f})}{n}$
$-\infty$	...	...	1.00000	...	...
...	...	...	...	0.00347	3.47
0.5	-10.158	-2.70	0.99653	...	...
...	...	...	...	0.01153	11.53
2.5	-8.158	-2.17	0.98500	...	...
...	...	...	...	0.03550	35.50
4.5	-6.158	-1.64	0.94950	...	...
...	...	...	...	0.08517	85.17
6.5	-4.158	-1.10	0.86433	...	...
...	...	...	...	0.14867	148.67
8.5	-2.158	-0.57	0.71566	...	...
...	...	...	...	0.19971	199.71
10.5	-0.158	-0.04	0.51595	...	...
...	...	...	...	0.20388	203.88
12.5	1.842	0.49	0.31207	...	...
...	...	...	...	0.15821	158.21
14.5	3.842	1.02	0.15386	...	...
...	...	...	...	0.09329	93.29
16.5	5.842	1.55	0.06057	...	...
...	...	...	...	0.04181	41.81
18.5	7.842	2.08	0.01876	...	...
...	...	...	...	0.01423	14.23
20.5	9.842	2.61	0.00453	...	...
...	...	...	...	0.00371	3.71
22.5	11.842	3.15	0.00082	...	...
...	...	...	...	0.00082	0.82
$\infty$	...	...	0.00000	...	...
<b>Total</b>				<b>1.00000</b>	<b>1000.00</b>

\*From Appendix 1.  
Source: Table A7.1.

From data of this nature a telephone company may budget in advance the cost of replacements of its poles and may purchase or contract for purchase upon the basis of these figures. A lumber dealer, long engaged in the sale of such poles, commented upon the possibility of using these fitted data as an indication of what might be expected of his salesmen.

Whether or not a normal curve is an appropriate type of function to fit to a frequency distribution may be ascertained in advance of the fit.

1. The appearance of a histogram of the data gives us a clue to the suitability of the normal curve. Although this is a crude guide, nevertheless it is possible to observe the presence of marked skewness or kurtosis. It has already been remarked that the appearance of the histogram of Chart A7.1 gives us little reason to doubt the suitability of the normal curve.

2. The data may be cumulated, as in Table A7.3, and percentage frequencies may then be calculated and plotted on probability paper as in Chart A7.2. An idea of the shape of a frequency distribution can often be obtained by examining the graph.

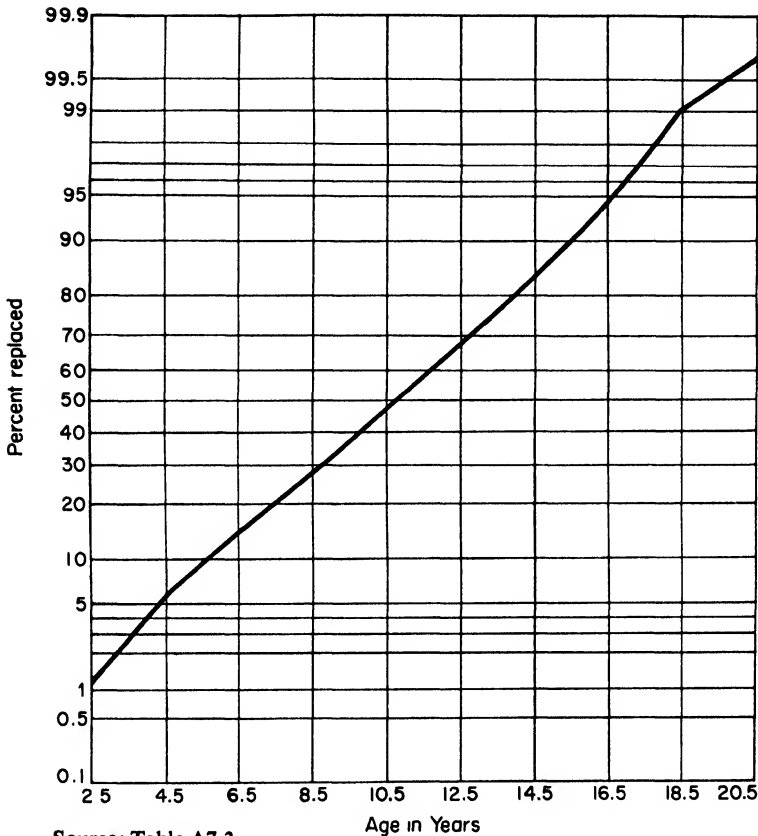
**TABLE A7.3: CUMULATIVE FREQUENCY DISTRIBUTION OF LIFE IN YEARS OF 1000 WOODEN TELEPHONE POLES**

<i>Class limit X</i>	<i>Number less than X</i>	<i>Percent less than X</i>
2.5	11	1.1
4.5	58	5.8
6.5	145	14.5
8.5	279	27.9
10.5	479	47.9
12.5	677	67.7
14.5	841	84.1
16.5	943	94.3
18.5	991	99.1
20.5	997	99.7
22.5	1000	100.0

Source: Table A7.1.

If the distribution is normal, the curve is a straight line. If the distribution is positively skewed, the curve tends to be concave from below at the lower end. If the distribution is negatively skewed, the curve tends to be concave from above at the upper end. If the distribution is leptokurtic, the curve tends to be S-shaped, flattening out at the ends. If the distribution is platykurtic, the curve tends to become steeper at the ends. In the present case the line is approximately straight, except at the ends. There seems to be little clear evidence from the curve as to the nature of the distribution, especially in view of the paucity of cases in the tails of the distribution, which causes those frequencies to be subject to considerable sampling error.

Note that the percentage for  $X = 22.5$  is not plotted, since normal

**CHART A7.2: LIFE EXPERIENCE OF WOODEN TELEPHONE POLES SHOWN ON PROBABILITY PAPER.**

Source: Table A7.3.

probability paper cannot extend to 100 percent. It is worth noting, also, that unequal class intervals do not invalidate the use of normal probability paper.

3. Measures of skewness and kurtosis may be computed and tested to ascertain if the amount of skewness and kurtosis present is significant. For the telephone data

$$a_3 = -0.063; \quad a_4 = 2.618; \quad a_4 - 3 = -0.382$$

Thus the frequency distribution is slightly skewed in a negative direction and is platykurtic, since  $a_3$  and  $a_4 - 3$  are zero for a normal population.

If we should take a large number of samples of  $n$  items each from a normal population and compute the value of  $a_3$  and  $a_4$ , we could form a frequency distribution of  $a_3$  values and another of  $a_4$  values. We could then compute the standard deviation of the  $a_3$  values and the  $a_4$  values. If the number of samples

was sufficiently large we would tend to get these results for large sample sizes:

$$\sigma_{a_3} \doteq \sqrt{\frac{6}{n}}; \quad \sigma_{a_4} \doteq \sqrt{\frac{24}{n}}$$

Although  $a_3$  is distributed symmetrically but not quite normally and the distribution of  $a_4$  has considerable positive skewness, it is nevertheless true that, if the population is normal, sample values of  $a_3$  and  $a_4 - 3$  are twice as large in absolute value as their standard errors or larger only (very roughly) about one time in twenty.

In the present instance

$$\sigma_{a_3} \doteq \sqrt{\frac{6}{1000}} = 0.077 \quad \text{and} \quad \frac{a_3}{\sigma_{a_3}} = \frac{-0.063}{0.077} = -0.82$$

so we accept the hypothesis that  $\alpha_3 = 0$  and conclude that the distribution is not significantly skewed. However,

$$\sigma_{a_4} \doteq \sqrt{\frac{24}{1000}} = 0.155 \quad \text{and} \quad \frac{a_4 - 3}{\sigma_{a_4}} = \frac{-0.382}{0.155} = -2.46$$

so we reject the hypothesis that  $\alpha_4 = 3$  and conclude that the distribution is significantly platykurtic.<sup>(1)</sup>

4. Whether or not a normal curve is an appropriate type of curve for use in describing a frequency distribution can also be decided after the expected frequencies have been computed and can be done by the chi-square test of "goodness of fit." The test is carried out in the manner described in Chapter 17 with the number of degrees of freedom equal to the number of classes minus 3, since the observed and fitted distributions were made to agree in three respects:  $\bar{X}$ ,  $(SD)^2$ , and  $n$ . The student can verify that at  $\alpha = 0.05$  the chi-square test does not discredit the hypothesis that the telephone pole distribution is normal.

As a final point, when the chi-square distribution is used to test goodness-of-fit, classes where  $\hat{f} < 1$  should be grouped with adjacent classes so that every class will contain an expected frequency of at least one. Such grouping is often necessary for the end classes in the distribution.

<sup>(1)</sup> More exact tests of significance of  $a_3$  and  $a_4$  can be made by use of Tables 34B and 34C of E. S. Pearson and H. O. Hartley (editors), *Biometrika Tables for Statisticians*, 3rd., ed., Vol. I, Cambridge University Press, Cambridge, pp. 207-208. On p. 68, Pearson and Hartley give an illustration of the use of these tables. In the *Biometrika* tables, the symbol  $\sqrt{b_1}$  is used instead of  $\sigma_{a_3}$  and  $b_2$  is used instead of  $\sigma_{a_4}$ .

# 8

## Introduction to Statistical Inference

In previous chapters we have pointed out that homogeneous statistical data are characterized by similarity and variability. There is a tendency for the data to be attracted toward some central value, but there is also a tendency for it to spread out in such a way that observations gradually become less frequent as we move away from the central value. We have also discussed various ways in which statistical data are summarized and have devoted considerable effort to the presentation of probability and probability distributions.

The time has come to tie these ideas together, for they are closely related. In statistical work we deal mainly with samples, but we are not so much interested in the sample as we are in the population from which the sample came. Thus in an examination an instructor asks a sample of questions, for he cannot examine a student on every point included in the course. But the examination is supposed to give the instructor some idea of what the student knows about the course as a whole. When we select a random sample of 46 corporation stocks and compare the average gain in price with that of a sample selected on some other basis by an investment counselor, we are not so much interested in the comparison of these two samples as we are in knowing what we can expect if we follow the counselor's advice. Are his selections really random ones after all, in spite of the ostensibly scientific method of selection? When we take samples of four pieces of cloth at hourly intervals and measure the breaking strength of each sample, we are interested in whether the process is being carried on under uniform conditions.

Thus, what we are mainly interested in is making inferences. We can never be sure that our inferences are correct because of the variability of statistical data and the resultant variability among the samples, but we can avoid being wrong too often and too much if we base our inferences on probability theory.

*By statistical inference is meant making a probability judgment concerning a population on the basis of one or more samples.*

*Statistics is mainly concerned with making statistical inferences as a basis for making decisions.* Thus, an instructor decides on the basis of examinations that a student has a satisfactory knowledge of the course, and he records a passing grade. An investor may decide, on the basis of a sample, that an investment analyst makes selections that are better than random, and so he subscribes to the service. A quality control engineer decides, after examining the control chart, that the process is "out of control." Therefore, he looks for the source of the trouble.

One point may disturb the reader. If one inspects the entire output of a factory during a given day, how can that be regarded as a sample? We might think of it as a sample of the output on the other days of the current year, but this would not be valid unless the same manufacturing conditions existed throughout the year. On the other hand, it could validly be regarded as a sample from an infinite theoretical population of units that would be produced by the same cause system. What you consider to be the population depends on your purpose.

Types of statistical inferences are generally classified as follows:

1. Making estimates.
  - a. Point estimates.
  - b. Interval estimates.
2. Testing hypotheses.

We shall discuss each of these briefly and as simply as we can without robbing the concepts of too much content.

Just as statistics describe a sample, so parameters describe a population. Often a statistic is considered to be an estimate of a parameter. We have discussed in past chapters the following statistics and parameters, which will be used in this chapter.

<i>Measure</i>	<i>Statistic</i>	<i>Parameter</i>
Mean	$\bar{X}$	$\mu$
Variance	$s^2$	$\sigma^2$
Standard deviation	$s$	$\sigma$
Number of defective items	$d$	$D$
Number of items	$n$	$N$
Proportion defective	$p = d/n$	$P = D/N$

## 8.1 POINT ESTIMATORS OF THE POPULATION MEAN AND VARIANCE

The most commonly used measure of population location is the population mean  $\mu$ , and the most commonly used measure of population dispersion is the population variance  $\sigma^2$ , or standard deviation  $\sigma$ .

**Finite and Infinite Populations.** The method of computing  $\mu$  and  $\sigma^2$  depends on whether the population is finite or infinite. Populations may be considered as being finite or infinite, depending upon the number of elements that they contain or the method of sampling used to investigate them. Technically a population consisting of all of the real numbers between zero and one has an infinite number of elements. Populations containing a hypothetically infinite number of elements are also often considered as infinite. An example is all of the output that could be imagined (almost imagined, at least) of a given process.

However, the number of elements contained in a population is not the only criterion used by statisticians to classify populations as finite or infinite, and a second one is associated with the method in which the population is sampled. If a population can be indefinitely sampled without altering its composition, the population is considered to be infinite. For example, let an urn contain four balls numbered 1, 2, 3, and 4. Now let us define two experiments.

1. Draw a ball from the urn, record its number, and discard the ball. This is sampling without replacement. Each time a ball is drawn, the composition of the population (contents of the urn) is altered.

2. Draw a ball from the urn, record its number, and return the ball to the urn. This is sampling with replacement. The composition of the population is not altered by the experiment, and the population is considered to be infinite even though it is composed of a finite number of elements.

**Estimating  $\mu$  and  $\sigma^2$ : Infinite Population.** Suppose that we take samples from the above population with replacement and with regard to order, of size  $n = 2$ . Table 8.1, also called a sampling matrix, shows the possible pairs we might draw. There are 16 such pairs, and, in general, when one is sampling distinct elements with replacement and with regard to order, there are  $N^n$  different ways of selecting samples of size  $n$  from a population of size  $N$ . The term "with regard to order" means that we consider samples such as 3, 1 and 1, 3 to be different even though they contain the same elements. Thus, we consider the order in which the sample was drawn to be important.

Since the population is considered to be infinite, we regard the elements of the population as having probabilities associated with them, and  $\mu$  and  $\sigma^2$  are

**TABLE 8.1: SAMPLING MATRIX FOR SAMPLES OF SIZE  $n = 2$ , DRAWN FROM A POPULATION OF SIZE  $N = 4$ , WITH REPLACEMENT AND WITH REGARD TO ORDER**

Item	1	2	3	4
1	1, 1	1, 2	1, 3	1, 4
2	2, 1	2, 2	2, 3	2, 4
3	3, 1	3, 2	3, 3	3, 4
4	4, 1	4, 2	4, 3	4, 4

defined respectively as the mean and variance of this probability distribution.

$$\mu = E(X) = \sum [X \cdot \text{Prob}(X)] \quad (8-1)$$

$$\sigma^2 = \sum \{[X - E(X)]^2 \cdot \text{Prob}(X)\} \quad (8-2)$$

Then, since each of the four elements in the population has an equal probability of occurring, 0.25, we have

$$\mu = 1(0.25) + 2(0.25) + 3(0.25) + 4(0.25) = 2.5$$

$$\begin{aligned} \sigma^2 &= (1 - 2.5)^2(0.25) + (2 - 2.5)^2(0.25) + (3 - 2.5)^2(0.25) + (4 - 2.5)^2(0.25) \\ &= 1.25 \end{aligned}$$

Our next task is to demonstrate that  $\bar{X}$  and  $s^2$  are unbiased estimators of  $\mu$  and  $\sigma^2$ , respectively, as we have previously asserted. Table 8.2 shows the results of calculating  $\bar{X}$  and  $s^2$  for each of the possible samples given in Table 8.1. For example, the sample containing the elements 1, 3 gives sample mean and variance

$$\bar{X} = \frac{1 + 3}{2} = 2$$

$$s^2 = \frac{(1 - 2)^2 + (3 - 2)^2}{1} = 2$$

Notice that there are seven different possible sample means and four different possible sample variances. Also shown in Table 8.2 are probabilities

**TABLE 8.2: POSSIBLE VALUES OF  $\bar{X}$  AND  $s^2$  AND ASSOCIATED PROBABILITIES AS CALCULATED FROM TABLE 8.1**a. FOR  $\bar{X}$ 

$\bar{X}$	$\text{Prob}(\bar{X})$	$\bar{X} \cdot \text{Prob}(\bar{X})$
1.0	1/16	1/16
1.5	2/16	3/16
2.0	3/16	6/16
2.5	4/16	10/16
3.0	3/16	9/16
3.5	2/16	7/16
4.0	1/16	4/16
Total	1.0	2.5

b. FOR  $s^2$ 

$s^2$	$\text{Prob}(s^2)$	$s^2 \cdot \text{Prob}(s^2)$
0	4/16	0/16
0.5	6/16	3/16
2.0	4/16	8/16
4.5	2/16	9/16
...	...	...
...	...	...
...	...	...
Total	1.0	1.25



associated with the various values of  $\bar{X}$  and  $s^2$ . These are the probabilities of making a simple random selection of a pair of balls that will yield the associated statistic. The student can verify that only one of the sixteen possible samples has a sample mean of one. Therefore, under simple random sampling, it has one chance in sixteen of being selected, since each of the possible samples have an equal probability of being selected. We assume simple random sampling throughout this chapter.

Having listed the possible sample statistics and their associated probabilities in Table 8.2, we now compute the expected value of  $\bar{X}$  and  $s^2$  in the usual manner.

$$E(\bar{X}) = \sum [\bar{X} \cdot \text{Prob}(\bar{X})] = \mu$$

$$E(s^2) = \sum [s^2 \cdot \text{Prob}(s^2)] = \sigma^2$$

We see from Table 8.2 (in the row labeled "Total") that the expected values of  $\bar{X}$  and  $s^2$  are respectively equal to  $\mu$  and  $\sigma^2$ . Thus, we have shown that  $\bar{X}$  and  $s^2$  are unbiased estimators of  $\mu$  and  $\sigma^2$ , respectively. If the symbol  $\theta$  is used to mean parameter, and  $\hat{\theta}$  to mean statistic, then

$$E(\hat{\theta}) = \theta$$

if a statistic is unbiased.

**Estimating  $\mu$  and  $\sigma^2$ : Finite Population.** Consider now the sampling matrix given in Table 8.3. It represents the possible samples that can be taken without replacement and without regard to order from our hypothetical population. There are no elements on the northwest-southeast diagonal of this matrix, since we are sampling without replacement; i.e., samples such as 1, 1 cannot be drawn. Also, we do not list elements below this diagonal, since including these elements in our calculations will not have any effect on the results. (The student may verify this statement in Problem 1.) Thus, we do *not* consider samples such as 3, 1 and 1, 3 to be different. This is the meaning of the term "without regard to order." Notice that there are six pairs listed and, in general, when one is sampling without replacement and without regard to order, there are  $\binom{N}{n}$  different ways of selecting

**TABLE 8.3: SAMPLING MATRIX FOR SAMPLES OF SIZE  $n = 2$ , DRAWN FROM A POPULATION OF SIZE  $N = 4$ , WITHOUT REPLACEMENT AND WITHOUT REGARD TO ORDER**

Item	1	2	3	4
1		1, 2	1, 3	1, 4
2			2, 3	2, 4
3				3, 4
4				

samples of size  $n$  from a population of size  $N$ . In this case

$$\binom{4}{2} = \frac{4!}{2!2!} = 6$$

Sampling in this way is often called *binomial sampling* and is very commonly done in statistics. Again we define the two parameters of interest,  $\mu$  and  $\sigma^2$ .

$$\mu = \frac{\sum [X \cdot f(X)]}{N} \quad (8-3)$$

$$\sigma^2 = \frac{\sum [(X - \mu)^2 \cdot f(X)]}{N - 1} \quad (8-4)$$

In computing  $\sigma^2$  we divide by  $N - 1$  instead of  $N$  because  $\mu$  is computed from the  $X$  values, which restricts the freedom of the  $X$  values to vary about  $\mu$ . If  $N = 1$ , there is no opportunity for  $X$  to differ from  $\mu$ ; if  $N = 2$ , each  $X$  value has  $\frac{1}{2}$  of an opportunity; if  $N = 3$ , each  $X$  value has  $\frac{2}{3}$  of an opportunity; and so on. In the previous case of an infinite population we could not, by definition, take cognizance of the infinite sample size. Also notice that we are using frequencies rather than probabilities in computing  $\mu$  and  $\sigma^2$ . Thus, in the present example, all  $X$  values have the same frequency, 1, and

$$\mu = \frac{1(1) + 2(1) + 3(1) + 4(1)}{4} = 2.5$$

$$\sigma^2 = \frac{(1 - 2.5)^2(1) + (2 - 2.5)^2(1) + (3 - 2.5)^2(1) + (4 - 2.5)^2(1)}{3} = 1.667$$

Again, we list the results of calculating  $\bar{X}$  and  $s^2$  for each of the possible samples. These calculations are summarized in Table 8.4, which is similar to

**TABLE 8.4: POSSIBLE VALUES OF  $\bar{X}$  AND  $s^2$  AND ASSOCIATED PROBABILITIES AS CALCULATED FROM TABLE 8.3**

a. FOR $\bar{X}$			b. FOR $s^2$		
$\bar{X}$	$Prob(\bar{X})$	$\bar{X} \cdot Prob(\bar{X})$	$s^2$	$Prob(s^2)$	$s^2 \cdot Prob(s^2)$
1.5	1/6	1.5/6	0.5	3/6	1.5/6
2.0	1/6	2.0/6	2.0	2/6	4.0/6
2.5	2/6	5.0/6	4.5	1/6	4.5/6
3.0	1/6	3.0/6	...	...	...
3.5	1/6	3.5/6	...	...	...
Total	1	2.5	Total	1	1.667

Table 8.2. Again, we assume that random sampling will insure that each of the six pairs will have the same probability of being drawn. It is clear that  $\bar{X}$  and  $s^2$  are still unbiased estimators of  $\mu$  and  $\sigma^2$  respectively since the mean values of their sampling distributions are the same as the values of the parameters.

## 8.2 POINT ESTIMATORS OF THE POPULATION PROPORTION DEFECTIVE

Let us compute the expected values of  $p$  and  $d$  for samples of 3 taken from an infinite population that is 40 percent defective ( $P = 0.4$ ). The probability distribution is given in Table 8.5 and was previously calculated using the binomial distribution in Table 6.4.

**TABLE 8.5: COMPUTATION OF  $E(p)$  AND  $E(d)$  FROM SAMPLES FROM AN INFINITE POPULATION WITH  $P = 0.4$**

$d$	$p$	$Prob(d) \text{ and } Prob(p)$	$p \cdot Prob(p)$	$d \cdot Prob(d)$
0	0.000	0.216	0.000	0.000
1	0.333	0.432	0.144	0.432
2	0.667	0.288	0.192	0.576
3	1.000	0.064	0.064	0.192
Total	...	1.000	0.400	1.200

Source: Table 6.4.

From Table 8.5 we see that

$$E(p) = P$$

or 0.4, and that

$$E(d) = nP$$

or  $3(0.4) = 1.2$ .

Therefore, we may say that  $p$  and  $d$  are unbiased estimators of  $P$  and  $nP$ , respectively. However, since we have used the binomial distribution we have assumed an infinite population. The case of a finite population will be covered in Chapter 12.

## 8.3 SOME QUALITIES OF A GOOD ESTIMATOR

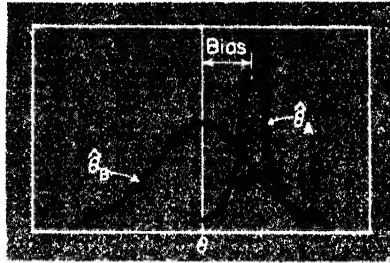
There are various qualities that are desirable in a statistic when it is taken as an estimator of a population parameter.

1. It should be unbiased.
2. It should be reliable.

**Lack of Bias.** We have already discussed lack of bias for several statistics; we review the results of this discussion below.

- $X$  is an unbiased estimator of  $\mu$ .
- $\bar{X}$  is an unbiased estimator of  $\mu$ .
- $s^2$  is an unbiased estimator of  $\sigma^2$ .
- $p$  is an unbiased estimator of  $P$ .
- $d$  is an unbiased estimator of  $nP$ .

**CHART 8.1: SAMPLING DISTRIBUTIONS OF TWO STATISTICS** ( $\hat{\theta}_A$  is precise but inaccurate.  $\hat{\theta}_B$  is accurate but imprecise.)



**Reliability.** The word “reliable” is somewhat vague, for it can have shades of meaning. A statistic is reliable if the variance of the statistic about its expected value is small; that is to say, if the estimator is *precise*. Thus

$$E[\hat{\theta} - E(\hat{\theta})]^2$$

is a measure of precision and  $\hat{\theta}$  is said to be precise if this quantity is a minimum. A statistic is also said to be reliable if the mean square deviation of the statistic about the parameter (mean square error) is small; that is to say, if the statistic is *accurate*. Thus

$$E[(\hat{\theta} - \theta)^2]$$

is a measure of accuracy and  $\hat{\theta}$  is said to be accurate if this quantity is a minimum. If the statistic is unbiased, then the distinction between precision and accuracy disappears, since

$$E(\hat{\theta}) = \theta$$

A good argument can be made for preferring a statistic that has smallest variance, even though it may be biased (provided that the bias is known). In Chart 8.1 the sampling distribution belonging to the statistic  $\hat{\theta}_A$  has a relatively small variance, but the statistic  $\hat{\theta}_A$  is a biased estimator of  $\theta$ . The statistic  $\hat{\theta}_B$  has no bias but has greater variance than does  $\hat{\theta}_A$ . It is often very difficult to choose the “best” estimator from a class of available estimators, and there is a vast literature on the choice of criteria. It can be argued, for example, that one should select from unbiased estimators that estimator for which the mean square deviation is the smallest. Such an estimator is called a minimum variance unbiased estimator.<sup>(1)</sup> All of the estimators discussed in this chapter are minimum variance unbiased estimators.

<sup>(1)</sup> Often the word *efficiency* is used in referring to reliability. A statistic is *asymptotically efficient* if, as the sample size approaches  $\infty$ , the statistic approaches the parameter, and the distribution of the statistic approaches the normal form with minimum variance. Efficiency is also used as a measure of reliability. When so used, it is either the ratio of the mean square deviations of two statistics when the sample sizes are the same, or the ratio of the sample sizes when the mean deviations are equal.

### 8.4 VARIANCE AND STANDARD ERROR OF THE MEAN

In the previous section we noted that the reliability of an unbiased statistic, such as  $\bar{X}$ , can be measured by its variance

$$\sigma_{\bar{X}}^2 = E(\bar{X} - \mu)^2 = \Sigma [\bar{X} - \mu]^2 \cdot \text{Prob}(\bar{X}) \quad (8-5)$$

The symbol  $\sigma_{\bar{X}}^2$  is called the variance of the sample mean. It is also true that the variance of the sample mean can be expressed as

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n} \left( 1 - \frac{n}{N} \right) \quad (8-6)$$

We now wish to demonstrate that Eq. (8-6) is true and to illustrate the meaning of the variance of a statistic.

**Infinite Population.** From Table 8.2 we see that the variance of the sample mean is

$$\sigma_{\bar{X}}^2 = (1.0 - 2.5)^2 \left( \frac{1}{16} \right) + (1.5 - 2.5)^2 \left( \frac{2}{16} \right) + \cdots + (4.0 - 2.5)^2 \left( \frac{1}{16} \right) = 0.625$$

which can be obtained directly by using Eq. (8-6)

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n} \left( 1 - \frac{n}{N} \right) = \frac{1.25}{2} \left( 1 - \frac{2}{\infty} \right) = 0.625$$

since the term  $(1 - 2/\infty) = 1$ . Because the term  $(1 - n/N)$  approaches 1 as  $N$  approaches  $\infty$ , it is sometimes called a *finite population correction factor*. In the case of very large populations, or in the case where the sample size is very small relative to the size of the population, the finite population correction factor can be safely ignored, and the “working” formula for the variance of the sample mean is

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n} \quad (8-7)$$

The square root of the variance of the sample mean is usually called *standard error of the mean* (sometimes the standard deviation of the sample mean).

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{1 - \frac{n}{N}} \quad (8-8)$$

For an infinite population

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \quad (8-9)$$

Notice carefully in Eq. (8-9) that the standard error of the mean varies inversely with the square root of  $n$ . In fact, in the limiting case when the

sample size is infinitely large, the standard error of the mean is zero, which means that there will be no variability in the distribution of sample means—the single sample mean is identical to the population mean, since  $\bar{X}$  is an unbiased estimator of  $\mu$ .

**Finite Population.** The student can verify from Table 8.4 that the variance of the sample mean is

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \left(1 - \frac{n}{N}\right) = \frac{1.667}{2} \left(1 - \frac{2}{4}\right) = 0.417$$

Notice again that when the sample size is the same as the population size,  $n = N$ , the variance of the sample mean is zero. Also notice that the finite population correction factor may be thought of as the proportion of the population *not* sampled. As before, the standard error of the mean is the square root of the variance of the sample mean.

## 8.5 VARIANCE AND STANDARD ERROR OF THE SAMPLE PROPORTION

For infinite populations the variance of the sample proportion is given by

$$\sigma_p^2 = E(p - P)^2 = \frac{PQ}{n} \quad (8-10)$$

as the student can verify by using Table 8.5. The student can also verify from the same table that the variance of the sample number defective is

$$\sigma_d^2 = E(d - nP)^2 = nPQ \quad (8-11)$$

The standard errors of the sample proportion and sample number defective, respectively, are found by extracting the square root of their variances. The case of finite populations will be covered in Chapter 12.

## 8.6 INTERVAL ESTIMATES

In the past few sections we saw that for a variety of statistics the expected value of the sampling distribution was the same as the value of the parameter. However, if we estimate a parameter using a statistic based upon a single sample, we can never be sure that the estimate of the parameter is correct even if the statistic is unbiased. We can only assert that if the process were repeated many times, the average value of the estimates would be the correct value of the parameter.

Since we can never be sure that an estimate of a parameter is correct, it is desirable that we state not only our best single estimate, or point estimate,

but also a pair of estimates, one above our point estimate and the other below it. This pair of estimates is made in such a way that it is a fair bet (at stated odds) that the parameter is between them. These estimates are also called confidence limits. If we wish to be very confident of our stated limits, we must make them rather wide, but if we are willing to take a big chance we can put them close together. How close together the limits are depends also on other factors. If there is not much variability in the sample, the limits will be closer together than if there is great variability. The limits will be closer together for a large sample than for a small sample. If a reliable statistic is used in determining the limits, they will be closer together than they will be if an unreliable one is used.

Suppose that we are interested in the average weight of the population of all male students at a certain university. On the basis of a random sample of size 100, we might calculate a sample mean  $\bar{X}$  of 165 pounds. In addition to making only the point estimate that the population mean is 165 pounds, we may, using methods to be discussed in a later chapter, make an interval estimate of the population mean. Such an estimate might be that the true population mean lies in the interval  $165 \pm 5$  pounds. The limits of the interval (confidence limits) are thus 160 pounds and 170 pounds. In addition, we may attach a degree of confidence to this estimate of, say, 95 percent. The 95 percent confidence interval says that if we were to draw all possible random samples of size 100 from this population, and for each of the sample means construct an interval of  $\bar{X} \pm 5$  pounds, 95 percent of these intervals would enclose the population mean. It should be clear that, other things remaining the same, we would be less confident of the interval  $\bar{X} \pm 0.5$  pounds and completely confident of the interval  $\bar{X} \pm \text{infinity}$ . In the last case we are completely confident that what we have said is true, since we have said nothing.

## 8.7 TESTS OF HYPOTHESES

A hypothesis is a statement concerning the population and is one that is open to doubt. We test the hypothesis and accept or reject it on the basis of probability theory. The hypothesis is tested by making use of a pre-defined *decision rule*, which is applied to sample data and which guides the experimenter in deciding whether to accept or reject the hypothesis on the basis of the *outcome* of the sample or samples drawn. Rejection of a hypothesis presumably leads to one kind of *consequence* and acceptance of a hypothesis leads to another kind of consequence.

As an illustration, let us consider a problem in quality control. Here the hypothesis might be that the process is in control; i.e., that the population mean does not differ from some predetermined standard value. Such a hypothesis is sometimes called a *null hypothesis*, denoted symbolically as  $H_0$ . It is a hypothesis of no difference. The *alternative hypothesis*,  $H_1$ , would be

that the process is out of control. We can test the null hypothesis, but we cannot test the alternative hypothesis in this case. To say merely that the process is out of control is not specific enough; such a hypothesis does not specify how far out of control the process is.

It should be clear that we can either accept or reject the null hypothesis. The decision may be correct in either of two ways:

1. We may accept the null hypothesis when it is true.
2. We may reject the null hypothesis when it is false.

The decision may be in error in two ways:<sup>(2)</sup>

1. We may reject a null hypothesis when it is true (type I error).
2. We may accept a null hypothesis when it is false (type II error).

In the next chapters we will consider the formulation and the testing of certain hypotheses. We will also relate hypothesis testing to point and interval estimation.

---

## PROBLEMS

1. Take samples of size  $n = 2$  from the population of four observations,

$X: 1, 3, 5, 6$

- i. Without replacement and without regard to order.
  - ii. Without replacement and with regard to order.
  - iii. With replacement and with regard to order.
- a. Show that  $\bar{X}$  and  $s^2$  are unbiased estimators of  $\mu$  and  $\sigma^2$ , respectively, under all three sampling plans.
  - b. Calculate  $\sigma_{\bar{X}}^2$  and  $\sigma_X^2$  under all three sampling plans.
2. Show, using Table 8.1 and Table 8.3, that  $E[(SD)^2] \neq \sigma^2$ . Can you correct for this bias using the results given in Chapter 4?
3. A statistic  $\hat{\theta}$  has the distribution given below. Calculate  $E(\hat{\theta})$  and  $\sigma_{\hat{\theta}}^2$ . If  $\theta = 2$ , is this statistic biased? Plot the distribution, indicating  $\theta$  and  $E(\hat{\theta})$ .

$\hat{\theta}$	Prob ( $\hat{\theta}$ )
1	0.2
2	0.5
3	0.2
4	0.1

---

<sup>(2)</sup> These errors are also called errors of the first and second kind.



4. A parameter  $\theta = 3$ . The distributions of two statistics are given below. Which statistic do you prefer? Why? Plot these distributions.

$\hat{\theta}_1$	$Prob(\hat{\theta}_1)$
1	0.2
3	0.6
5	0.2

$\hat{\theta}_2$	$Prob(\hat{\theta}_2)$
1	0.1
2	0.2
3	0.4
4	0.2
5	0.1

5. If a population is continuous and we can attach a degree of confidence of 100 percent to the interval  $\hat{\theta} \pm \infty$ , how much confidence can we attach to the interval  $\hat{\theta} \pm 0$ ? Explain what implications your answer has for point estimates under these conditions.

6. Each morning you formulate the null hypothesis: "It will not rain today." On days that you reject this null hypothesis you take a raincoat to work. On days that you accept this null hypothesis you do not take a raincoat. You do not wish to take a raincoat on days when there is no rain, nor do you wish to leave your raincoat at home on days when there is rain. You are perfectly happy if you wear your raincoat on rainy days and leave it at home on days when there is no rain.

- a. Formulate the two types of errors that you may make in testing your null hypothesis.
- b. Suppose we call  $\alpha$  (alpha) the probability of making a type I error.
  - i. What would  $\alpha$  be if you decided always to take your raincoat?
  - ii. What would  $\alpha$  be if you decided never to take your raincoat?

# 9

## Sampling Design

This chapter extends the ideas presented in the last chapter, and the methods to be presented apply to all types of investigations requiring the use of samples, such as industrial experiments or surveys of populations. A large proportion of the discussion, however, is directed toward survey sampling. This chapter may be omitted on first reading without destroying the continuity of presentation.

### 9.1 SOME BASIC IDEAS

**Factors Affecting Sampling Design.** In deciding upon the method of sampling and the design of a sample survey, several factors should be considered.

1. *Whether the population is homogeneous.* Data are said to be homogeneous (alike qualitatively) if all the units in the population are governed by the same set of causes. Often it is possible to divide the population into rational groups, among which different sets of causes are at work, but within each of which a constant set of causes is operating. In other words, a heterogeneous population can often be classified into homogeneous subgroups.

2. *The degree of precision required.* Precision refers to the uniformity of results that are to be expected from repeated samples of the same size and type from the same population. As stated in Chapter 8, reliability of an unbiased statistical measure, such as the arithmetic mean, is measured by its variance, which for randomly selected sample means is

$$\sigma_x^2 = \frac{\sigma^2}{n} \left( 1 - \frac{n}{N} \right)$$

or its standard error  $\sigma_x$ , which is the square root of the variance. If the standard error of the mean is used as the measure of reliability, then the reliability of arithmetic means computed from random samples of size  $n$  from a large population varies with the square root of  $n$ .

$$\sigma_x = \frac{\sigma}{\sqrt{n}}$$

Thus, other things being the same, the arithmetic mean of a sample of 10,000 items is 10 times as reliable as the arithmetic mean of a sample of 100 items. Large size alone, however, is no guarantee that a sample is reliable; reliability depends also on the method of sampling and the sampling design used (as well as on the accuracy of the observations).

3. *The cost of the sampling plan.* Although it is possible to increase the reliability of a sample by increasing its size, this procedure may not be economical. The cost may increase more or less proportionately with the sample size, whereas the reliability of the measures we are interested in computing from the sample increases much less rapidly. Other points, also, should be considered. A sample of 500 items comprised of five observations from each of 100 areas taken at random will cost more to collect than will a sample of 500 items comprised of 100 observations from each of five areas taken at random. The cost of sampling, including the cost of administering the plan, as compared with the cost of measurement, is sometimes the decisive consideration in testing materials. It should be noted that measurement may be destructive. If this is the case, design of the entire experiment, including the choice of the test statistic, should be such that the number of units measured be at a minimum.

**Definition of Population to be Sampled.** If one is sampling the farm population of a state, he must be able to distinguish between a farm, a garden, and a rural residence. Or, in an enumeration of industrial establishments, one might have to decide whether to try to include, for instance, a college student who strings tennis racquets occasionally in the evenings. Sometimes a cutoff point is established, such as sampling only firms making goods worth more than \$5000 during the year. The use of an incomplete list enables one to omit many small firms, thus saving time and money. Obviously, it is dangerous to extend conclusions beyond the population sampled.

**Choice of Sampling Units.** In a survey of buying habits should the ultimate sampling unit be the individual or the family? In a study of soap consumption the family would generally be selected as the sampling unit, because it is impossible to measure the separate consumption of the different members of a family. Often, however, the choice of sampling units

hinges on the reliability that will be obtained from different units compared with the cost of collecting the data.

**The Place for Exercise of Judgment.** Sampling may be either (1) probability sampling or (2) judgment sampling. This chapter is concerned with probability sampling. Although judgment is exercised in designing a probability sampling plan, the enumerator does not exercise judgment with respect to the units included in his sample. The chief advantages of probability sampling are twofold:

1. The operation of chance is likely to result in a more representative sample than is the exercise of judgment.
2. Random sampling methods result in a probability distribution and make it possible to estimate the magnitude of the sampling error.

If the sampler is permitted to decide which items to include in his sample, we have judgment sampling. Although it is possible by the exercise of good judgment to obtain a representative sample, usually this will not result. In an endeavor to obtain a reasonable average, some samplers will select only items that they consider typical. This selection will result in a sample that is more nearly uniform than the population. Other samplers may select small, medium, and large items in an endeavor to obtain representatives of all magnitudes, but such a procedure rarely results in the different sizes being represented in the correct proportion. In interviewing people, there is the danger that the interviewer will select for his sample mainly those who are pleasant to interview. Such procedures are usually even less likely to result in representative samples than is a conscious effort to select the best sample.

**Absolute and Relative Sample Size.** The question is often asked whether a sample that is 10 percent of the population is satisfactory; such a question cannot be answered. The standard error of a mean depends on both the absolute sample size and the relative sample size  $n/N$ . For a finite population we know that

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}$$

Suppose the standard deviation of the population is  $\sigma = 100$ . Let us see what happens to  $\sigma_{\bar{x}}$  as we vary  $n$  and  $N$ , using the data on page 116.

It is apparent that when the sample size is not more than 10 percent of the population, it does not make much difference what the relative size of the sample is. But the absolute sample size greatly affects the reliability of the sample.

$n = 25$ $N = 2500$ $\frac{n}{N} = 0.01$ $\sigma_{\bar{x}} = \frac{100}{5} \sqrt{0.99} = 19.9$	$n = 25$ $N = 500$ $\frac{n}{N} = 0.05$ $\sigma_{\bar{x}} = \frac{100}{5} \sqrt{0.95} = 19.5$
$n = 100$ $N = 2000$ $\frac{n}{N} = 0.05$ $\sigma_{\bar{x}} = \frac{100}{10} \sqrt{0.95} = 9.8$	$n = 100$ $N = 1000$ $\frac{n}{N} = 0.10$ $\sigma_{\bar{x}} = \frac{100}{10} \sqrt{0.90} = 9.5$

## 9.2 METHODS OF SAMPLING

There are two methods of sampling: random and systematic. All sampling designs utilize one or both of these methods.

**Random Sampling.** *Sampling is said to be random if each possible sample (combination of a given number of items) has the same probability of being drawn.* Suppose we have the presumably homogeneous population shown below, and we wish to obtain a random sample of three units from this population. What are the different combinations of three items, each of which has the same chance of being drawn?

Unit	Diameter (millimeters)
A	3
B	4
C	4
D	5
E	5
F	6

The student should not conclude that these data are necessarily heterogeneous because the different units are of a different size; it must be remembered that the distinction between homogeneity and heterogeneity is a qualitative, not a quantitative, distinction. Also, the student should not conclude that one is likely to want to take a sample of 3 from a population of 6. He is more likely to require a sample of 500 from a population of 100,000. But the above data provide an illustration of manageable proportions.

There are  $\binom{6}{3} = 20$  different samples of 3 that can be drawn from the above population of 6, without replacement and without regard to order.

These are shown below, together with the mean value of each sample, rounded to two decimal places.

<i>Sample</i>	<i>Mean</i>	<i>Sample</i>	<i>Mean</i>
<i>ABC</i>	3.67	<i>BCD</i>	4.33
<i>ABD</i>	4.00	<i>BCE</i>	4.33
<i>ABE</i>	4.00	<i>BCF</i>	4.67
<i>ABF</i>	4.33	<i>BDE</i>	4.67
<i>ACD</i>	4.00	<i>BDF</i>	5.00
<i>ACE</i>	4.00	<i>BEF</i>	5.00
<i>ACF</i>	4.33	<i>CDE</i>	4.67
<i>ADE</i>	4.33	<i>CDF</i>	5.00
<i>ADF</i>	4.67	<i>CEF</i>	5.00
<i>AEF</i>	4.67	<i>DEF</i>	5.33

A random sample is an appropriate type of sample for a homogeneous population. Samples of cord to be used in tires and to be tested for flex life and elongation may be selected at random, as may bolts (all of which are of the same material and made by the same or similar machines) to be tested for tensile strength, and likewise many other manufactured items.

When a random sample seems appropriate, one must grapple with the problem of achieving randomness. The following two procedures, which are called random sampling techniques, help attack this problem.<sup>(1)</sup>

1. *Thorough mixing.* One method is to mix the units together thoroughly and then draw the units for the sample in some unbiased manner. Sometimes the product being sampled is of such a character that it is very difficult to mix thoroughly. For instance, the heavy units may gravitate to the bottom, and a random sample will not be obtained by ordinary mixing methods. In other cases, the object under consideration is so bulky, fragile, or immobile that physical mixing is out of the question. Even under the most favorable circumstances, it is difficult to know when the units have been mixed *thoroughly enough*.

2. *Use of random numbers.* A method of overcoming this difficulty is to number the items and then select the desired number of units by use of a table of random numbers.<sup>(2)</sup> Assume that the following is a partial table of random numbers.

4	1	9	2	0
9	6	9	7	4
2	0	0	7	9
4	5	8	4	7
3	8	4	0	1

<sup>(1)</sup> A strong argument may be made that although there exist random sampling techniques, there cannot exist truly random sampling.

<sup>(2)</sup> A table of random numbers is given in Appendix 12.

If we number the units of our previous illustration consecutively from  $A = 1$  to  $F = 6$  and proceed with our random numbers from left to right, beginning with the first row, the random sampling numbers used are 4, 1, 2, and the sample is *DAB*. We omit the number 9, since no population element possesses this number. If we proceed from top to bottom, beginning with the column on the left, the random sampling numbers (ignoring duplicates, since we are sampling without replacement) are 4, 2, 3, and the sample is *DBC*. Random numbers may be used in any methodical manner decided upon before noticing the arrangement of items being sampled. Since we are assuming random sampling, each of the 20 possible samples has an equal probability of being drawn, and we may put the list of 20 sample means in the form of a probability distribution.

$\bar{X}$	<i>Prob</i> ( $\bar{X}$ )
3.67	0.05
4.00	0.20
4.33	0.25
4.67	0.25
5.00	0.20
5.33	0.05

For the purpose of future comparison we compute the variance of these sample means.<sup>(3)</sup>

$$\begin{aligned}
 \sigma_{\bar{X}}^2 &= E(\bar{X} - \mu)^2 = \sum [(\bar{X} - \mu)^2 \cdot \text{Prob}(\bar{X})] \\
 &= (3.67 - 4.50)^2(0.05) + (4.00 - 4.50)^2(0.20) + (4.33 - 4.50)^2(0.25) \\
 &\quad + (4.67 - 4.50)^2(0.25) + (5.00 - 4.50)^2(0.20) + (5.33 - 4.50)^2(0.05) \\
 &= 0.183
 \end{aligned}$$

**Systematic Sampling.** In systematic sampling we select units from the population at uniform intervals of time, space, or order of occurrence. Thus, if we wish to select a systematic sample of 3 from our hypothetical population of 6, there are only two possible samples.

<i>Sample</i>	<i>Mean</i>
<i>ACE</i>	4.0
<i>BDF</i>	5.0

<sup>(3)</sup> This could have been calculated by using

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n} \left( 1 - \frac{n}{N} \right)$$

as the student may verify.

The item number with which to start should be determined at random. Thus, using our abbreviated table of random numbers, and proceeding from left to right with the first row, we encounter digit 1 before we come to digit 2. The sample is therefore *ACE*.

It may be appropriate to use a systematic sample where the units are arranged in some more or less systematic manner. For instance, it would be convenient, and probably satisfactory, to select names from a telephone directory in this manner (if telephone subscribers constitute the population we wish to sample). Another illustration of a systematic sample is the drawing of such items as bolts or rivets from the production line. For a strictly systematic sample, the units would be selected singly and spaced at equal intervals of time or order of production.

A substantial advantage of systematic sampling is its low cost, which stems from its simplicity. A list of persons to be interviewed need not be obtained, and a sample of persons laboriously drawn. The interviewer need only to ring the bell at, say, every twentieth household as he follows a predetermined path. It is usually best to take a serpentine route, rather than working one street the full length from east to west, and then starting the next street at the east end.

Whether a systematic sample is more reliable than a random sample depends on the arrangement of items in the sampled population. If the values are arranged in exact order of magnitude, the variance of the mean for systematic sampling will not necessarily be smaller than for random sampling. For the example of this section the variance is:

$$\begin{aligned}\sigma_{\bar{X}}^2 &= E(\bar{X} - \mu)^2 = \sum [(\bar{X} - \mu)^2 \cdot \text{Prob}(\bar{X})] \\ &= (4 - 4.5)^2(0.5) + (5 - 4.5)^2(0.5) = 0.25\end{aligned}$$

which is larger than the variance computed for random sampling. If the items in the population tend to come in approximately periodic waves, so that, say, every twentieth item is large, then a 5 percent systematic sample might contain only the smallest items or only the largest items.

If a heterogeneous population is arranged into homogeneous groups, with random arrangement within each group, a systematic sample will be more reliable than a random sample because it assures that items of different magnitude are included in approximately the correct proportion. With a random sample the representation is a matter of chance.<sup>(4)</sup>

### 9.3 SAMPLING DESIGNS

**Simple Sampling.** The sampling designs that we have considered are *simple* designs: simple random sampling and simple systematic sampling.

<sup>(4)</sup> If we could be assured that the items are arranged in this manner, systematic sampling is identical with proportionately stratified random sampling.



Many other designs are available. In each case the object is to obtain unbiased results of a given degree of reliability: (1) with the minimum sample size, or (2) with the minimum expenditure of money. Or, stated conversely, with a given sample size or a given expenditure, we want to obtain results that are as reliable as possible (minimum variance among the means or other statistical measures).

**Stratified Sampling.** When heterogeneity is present in a population, it is usually desirable to classify the population into strata and select a random or systematic sample from each stratum. If the population can be split into perfectly homogeneous strata (each with a variance of zero), stratified sampling will afford the smallest possible sample variance.<sup>(6)</sup> Referring again to our hypothetical data and assuming that items *A*, *B*, *C*, and *D* constitute a product from the day shift, whereas *E* and *F* constitute a product from the night shift, we shall enumerate the twelve possible stratified random samples of 3, each of which includes 2 items from the day shift and 1 from the night shift.

<i>Sample</i>	<i>Mean</i>	<i>Sample</i>	<i>Mean</i>
<i>ABE</i>	4.00	<i>ABF</i>	4.33
<i>ACE</i>	4.00	<i>ACF</i>	4.33
<i>ADE</i>	4.33	<i>ADF</i>	4.67
<i>BCE</i>	4.33	<i>BCF</i>	4.67
<i>BDE</i>	4.67	<i>BDF</i>	5.00
<i>CDE</i>	4.67	<i>CDF</i>	5.00

Notice that each of these samples has the same chance of being drawn, but that there are 8 random samples that cannot be included among our stratified samples. Also notice that the means of the different samples tend to cluster more closely together for our stratified sampling method than for our random sampling method. This clustering can be seen more clearly if we form a probability distribution of the means.

$\bar{X}$	<i>Prob</i> ( $\bar{X}$ )
4.00	0.167
4.33	0.333
4.67	0.333
5.00	0.167
Total	1.000

<sup>(6)</sup> Often a proportionately stratified sample is used, but sometimes the proportion varies from stratum to stratum. The variance of the mean will be minimized if the number of sample units allocated to each stratum is proportional to the product of the size of the stratum and the standard deviation of the stratum. This method of allocation is sometimes referred to as the method of optimum allocation. See William G. Cochran, *Sampling Techniques* (New York: John Wiley and Sons, Inc., 1953), Sec. 5.5.

The variance of the means is

$$\sigma_{\bar{x}}^2 = (4.00 - 4.50)^2(0.167) + (4.33 - 4.50)^2(0.333) + (4.67 - 4.50)^2(0.333) + (5.00 - 4.50)^2(0.167) = 0.102$$

This value is much smaller than the variance of the simple random sample means, which was 0.183, because of greater homogeneity within the strata than for the population as a whole.

It should be obvious that the strata selected should be germane to the problem; otherwise there will be no increase in the reliability of the sample as compared with a simple random sample. For example, a study of expenditures of male white college seniors might not be improved by using as two strata, those with light hair and those with dark hair, but it doubtless would be helpful to separate fraternity and nonfraternity men. If the college seniors were to be stratified according to whether or not they are members of a fraternity and also according to academic standing (such as A, B, C, and D average grade), cross-classification is involved, and there are eight strata. If many strata, substrata, and sub-substrata are used, stratification becomes unwieldy.

Ordinarily the sample within each stratum should be taken at random, though occasionally a systematic sample may be appropriate. If enumerators are allowed to exercise their own initiative in selecting the items (sometimes called the "quota" method) within a stratum, they often select the most readily available cases, fail to check back on "not at home" calls, and do other things that result in a sample that is not representative.

One difficulty with stratified sampling is that at least some knowledge of the population must be available before strata can be properly selected. When a study is first undertaken, one's notion of what may constitute pertinent bases of classification may be based on meager information. It is therefore frequently advisable to conduct a *pilot study*, using a relatively small sample. Such a study is helpful in various ways.

1. It provides an estimate of the means of the different strata, thus giving an indication of what classifications are worth using.
2. It provides an estimate of the number of units in and the variances of the different strata, thus helping one to decide the optimum allocation of the sample among the different strata.
3. It provides an estimate of the response rates for the different strata, thus providing a basis for correction of the bias that arises from the differential response rate (especially when the questionnaire method is used).
4. It indicates the effectiveness of the different questions on the schedule.

**Cluster Sampling.** In sampling a manufacturing process there are at least three possible procedures.

1. Inspect every  $i$ th item or one item every  $j$  minutes. This is systematic sampling.
2. Inspect a random sample of  $n$  items from the output during each hour. This is stratified random sampling.
3. Inspect  $n$  consecutive items at intervals of approximately one hour. This is cluster sampling and is the method usually followed in process control.

Cluster sampling usually requires a larger sample to attain the same degree of reliability than does simple random sampling, because the different observations in the same cluster usually have about the same values.

An example of cluster sampling is an investigation where the ultimate sampling units are restaurants. It would be theoretically possible to study the operation of a random sample of restaurants covering the entire United States. If one were to spend a month investigating each restaurant, this might be the best procedure. If, however, the investigation involves only a 5-minute interview with each restaurant manager, such a procedure would quite likely involve prohibitive costs. Too much time would be devoted to traveling from place to place, and too little time would be spent in interviewing. It would be cheaper to make a random selection of *cities* and study each restaurant in the cities selected. We could perhaps reduce the proportion of the total cost devoted to traveling by selecting a sample of *states* at random and investigating each restaurant in the states selected. But although this might be cheaper per sampling unit, it might increase the sampling error too much.

The size of the individual clusters relative to the number of clusters selected obviously depends on:

1. The cost of alternative sampling plans. When the sampling is spread over a geographical area, the decisive factor is the cost of measurement or interviewing relative to the cost of traveling. In general, for a given sample size, it is cheaper to use a small number of large clusters than a large number of small clusters.
2. The reliability of the results obtained by the alternative plans. In general, for a given sample size, more reliable results are obtained from a large number of small clusters than from a small number of large clusters.

**Multistage Sampling.** If the units being studied are scattered over a geographical area, it is often economical to divide the territory into regions called *primary sampling units*, a number of which are selected either systematically or at random. Then from each of the regions that were selected by the sampling process, a number of subregions are selected systematically or at random. If each of the *ultimate sampling units* (which may be persons, households, farms, etc.) in a subregion is investigated, we have two-stage sampling, and the ultimate sampling units in a subregion constitute a cluster. If we select systematically or at random a sample from each subregion, we have three-stage sampling. Multistage sampling can be applied in many fields. For example, every hundredth box can be taken from the production line and samples taken of the contents of the boxes selected. The advantage of multistage sampling is that a larger number of units can be sampled than by use of a simple sampling design at the same cost.

**Area Sampling.** This is not really a separate type of design but is a term used to refer to sampling designs in which the primary sampling units are land areas. Usually area sampling utilizes both multistage sampling and cluster sampling.

**Sampling with Probability Proportional to Size.** In Table 9.1 are listed, in rank order, the cities in the United States with population of 700,000 or more in 1960. We wish to select a sample of ten department stores from these twelve cities for an intensive investigation.

**TABLE 9.1: CITIES IN THE UNITED STATES WITH POPULATION OF 700,000 OR MORE, AND SAMPLE OF TEN WITH PROBABILITY PROPORTIONAL TO SIZE**

<i>City</i>	<i>Population (in thousands)</i>	<i>Cumulative total (in thousands)</i>	<i>Sample</i>
1. New York	7782	7,782	1581; 3094; 6227
2. Chicago	3550	11,332	8550; 10,873
3. Los Angeles	2479	13,811	13,196
4. Philadelphia	2003	15,814	15,519
5. Detroit	1670	17,484	
6. Baltimore	939	18,423	17,842
7. Houston	938	19,361	
8. Cleveland	876	20,237	20,165
9. Washington, D.C.	764	21,001	
10. St. Louis	750	21,751	
11. Milwaukee	741	22,492	22,488
12. San Francisco	740	23,232	

*Source: U.S. Department of Commerce, Statistical Abstract of the United States, 1965, pp. 19-20.*

First, we notice that the total population (in thousands) of these twelve cities is 23,232. Since  $23,232/10 = 2323$ , a sampling interval of 2323 is appropriate.

Next, we select at random some 4-digit number between 1 and 2323. The first number in the table of random numbers given in Appendix 12 is 1581.

Starting with 1581, we add successive increments of 2323 until we have obtained 10 numbers. These are shown in the last column of Table 9.1 and indicate that three stores should be selected in New York, two in Chicago, and one each in five other cities.

Note that this method of sampling does not assure that New York will be represented by three stores, nor does it preclude Chicago from having three stores, nor does it preclude Houston from having one store. How it works depends on our random start. It would be just as satisfactory if the cities were arranged alphabetically, or at random, providing we used a random start.

**Random-point Sampling.** This method consists first of locating many points at random on a map. A cluster consisting of a given number of items nearest to each point is then included to form the sample. Such a procedure cannot usually be recommended for selecting a sample of farms, for example, since the random points are more likely to fall on or near large farms than small farms. This method, obviously, can be used only for geographical series. It might be worth considering for selecting a sample of land areas classified by use or erosion status, where the unit is not a farm but an acre or square mile, and for other similar purposes.

**Sequential Sampling.** It is sometimes expensive, and occasionally destructive, to test a raw material or manufactured product. In such instances it is desirable to draw inferences by testing a relatively small number of items. With double sampling, a small sample is first tested. If the sample is very good, the lot is accepted; if the sample is very bad, the lot is rejected; if it is intermediate, a second sample is taken and the product is accepted or rejected on the basis of the two samples combined. Multiple sequential sampling employs the same principle, but the decision to accept or reject will not necessarily be made until some predetermined number of samples is taken. With item-by-item sequential sampling, additional items are tested, but it may necessitate selecting a relatively large sample from which subsamples are taken. It is also rather expensive to administer because of the clerical work involved.

**Latin Square.** In agricultural work it is costly to perform an experiment, and it is important to have an efficient experimental design. Suppose we are considering the application of four types of treatment to four varieties of plant with four different types of soil. If we used a *factorial* design, which considers all of the possible combinations of factors, we would need a minimum of 64 observations.

By using a *Latin square* design we can get along with 16 observations. Let us arrange the varieties in columns and the treatments in rows and designate the soil conditions by the letters *A, B, C, D*. Then one Latin square is

TREATMENTS	VARIETIES			
	1	2	3	4
1	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>
2	<i>B</i>	<i>A</i>	<i>D</i>	<i>C</i>
3	<i>C</i>	<i>D</i>	<i>A</i>	<i>B</i>
4	<i>D</i>	<i>C</i>	<i>B</i>	<i>A</i>

Thus the third row of the second column shows the result of treatment 3 applied to variety 2, planted in soil *D*. In a Latin square each letter occurs once and only once in each column and in each row. Obviously there is more than one Latin square with 4 columns and 4 rows,<sup>(6)</sup> and the square or squares to be used can be selected at random. By averaging the values in a given column one obtains an unbiased estimate of that variety mean; by averaging the values in a given row one obtains an unbiased estimate of that treatment mean; by averaging the values of a given letter one obtains an unbiased estimate of the soil mean.

This type of design is applicable to industrial experimentation. For example, the factors may be machines, materials, and men. Or it can be used in sampling human populations. For example, cities could be selected for study of buying habits of wage earners after classifying the cities on the basis of size of city, mean temperature, and per capita income.

The Latin square is only one of many designs, some very complex, devised for the purpose of obtaining the maximum information at the least expense.

---

## PROBLEMS

### 1. Given the population

$$X: 1, 4, 1, 1, 4, 1$$

and the assumption that sampling is done without replacement and without regard to order:

- a. List all possible simple random samples of size  $n = 3$  that may be selected and use this list to compute  $\sigma_x^2$ .*
- b. List all possible simple systematic samples of size  $n = 3$  that may be selected and use this list to compute  $\sigma_x^2$ .*
- c. Stratify the population as follows:*

$$1, 1, 1, 1 \quad \text{and} \quad 4, 4$$

*and list all stratified samples of size  $n = 3$  that can be selected where two units are selected from the first strata and one unit is selected from the second strata. Use this list to compute  $\sigma_x^2$ .*

---

<sup>6</sup> For a tabulation of Latin squares see R. A. Fisher and F. Yates, *Statistical Tables for Biological, Agricultural and Medical Research*. (Edinburgh: 1949), Oliver and Boyd, Ltd., Table XV.

*Comment on any differences in reliability shown by your three sampling plans.*

**2. Define:**

- a. Cluster sampling.*
- b. Multistage sampling.*
- c. Area sampling.*
- d. Sampling with probability proportional to size.*
- e. Random point sampling.*
- f. Sequential sampling.*

**3. Write a  $5 \times 5$  Latin square.**

# 10

## Tests of Hypotheses and Confidence Limits for the Arithmetic Mean: Population Variance Known or Specified

In this chapter we formalize some of the ideas presented verbally in Chapter 8. Of the numerous statistics discussed up to this point we will, for the moment, consider only one—the arithmetic mean. We will also confine ourselves to a simple type of hypothesis where the population variance (or standard deviation) is either known or specified. Although such hypotheses are not the most common in practical statistical work, they do form a basis for the discussion of the case where the population variance is unspecified. Also, in quality control the population variance is often specified; thus, the discussion has some practical significance.

### 10.1 TWO-SIDED TEST

For many productive processes it is equally bad if the mean of the process deviates in either direction from some accepted norm. For example, it is equally bad if a bolt blank is either longer or shorter than some specified length. Consider the following decision problem of an engineer who is in charge of finishing bolt blanks produced in another department of his plant. The finishing department receives a very large lot of bolt blanks periodically from the extruding department. From past history it is known that these incoming blanks are distributed with a population variance  $\sigma^2$  of



400 (population standard deviation of 20 mm). It is desirable from the point of view of the finishing department that the average length of the blanks in any given lot be 200 mm. Each time a given lot of blanks is received by the finishing department, a decision must be made as to whether the lot should be accepted for finishing or rejected as being of unacceptable quality. If the average length of the blanks in a given lot is determined to be 200 mm, the lot will be accepted; if it is determined beyond reasonable doubt to be different from 200 mm, the lot will be rejected. Symbolically, the null and the alternative hypotheses (or family of hypotheses) are

$$H_0: \mu = 200 \text{ mm}$$

$$H_1: \mu \neq 200 \text{ mm}$$

and we will refer to 200 mm as  $\mu_0$ .

Assume for the time being that plant policy states that the decision to reject or accept a lot is to be made on the basis of a single random sample of size  $n = 100$  taken from the lot.

Given the sampling plan, the engineer must now formulate a decision rule. He knows that for large simple random samples, taken from a population with mean  $\mu_0$ , the distribution of sample means will be approximately normal with mean  $\mu_0$ . Therefore, he knows that sample means that are "greatly different" from  $\mu_0$  are less probable than sample means that are "not greatly different" from  $\mu_0$ . The term "greatly different," of course, depends upon the variation in the distribution of the sample means as measured by  $\sigma_{\bar{x}}$ . Thus, it would seem reasonable to formulate a decision rule that would lead to the rejection of  $H_0$  if a sample mean that is greatly different from  $\mu_0 = 200$  mm is encountered and acceptance of  $H_0$  if one not greatly different from  $\mu_0$  is encountered. Of course, he must include in his decision rule a statement of how large a difference is required for rejection of  $H_0$ . This statement must be made prior to the test, and its formulation will be discussed in Sec. 10.5. For the present, let us say that plant policy is that not more than 5 percent of all lots with average length of 200 mm be rejected by the finishing department. That is to say, the null hypothesis should be rejected only if the observed difference between  $\bar{X}$  and  $\mu_0$  is so large that the probability of a deviation as large or larger than the one observed is equal to or less than 0.05. We will call this *error probability*  $\alpha$  (alpha). It is the probability of rejecting  $H_0$  given that  $H_0$  is true (a type I error). This error probability is also called a *level of significance*.

Suppose that for a particular lot the engineer finds that  $\bar{X} = 205$  mm. Should he reject this lot? Since we are working with a very large population, the standard error of the mean is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

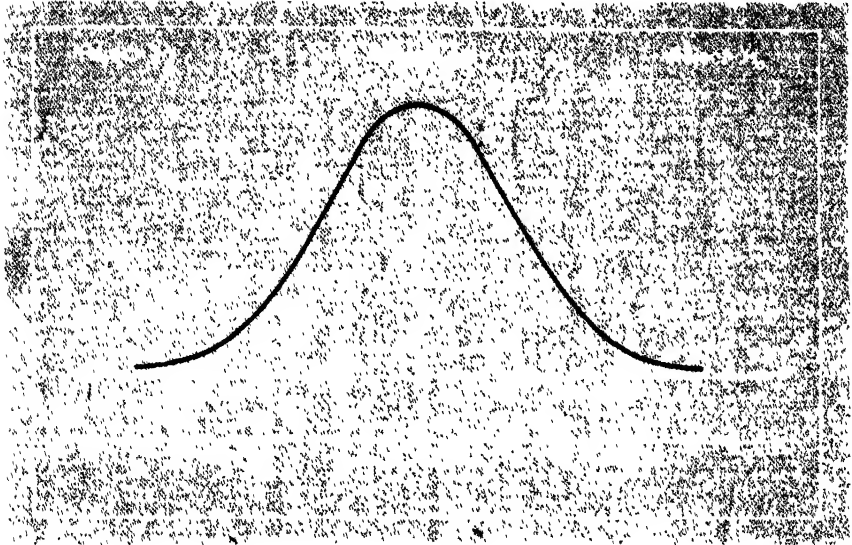
or in this case

$$\sigma_{\bar{x}} = \frac{20}{\sqrt{100}} = 2$$

Now the sample mean differs from  $\mu_0$  by 5 mm. In terms of the distribution of the sample means, is this deviation so large that the probability of a deviation this large or larger is equal to or less than 0.05?

Chart 10.1 shows the distribution of means of samples of  $n = 100$  taken at random from a very large lot with  $\sigma = 20$  mm and  $\mu_0 = 200$  mm. Notice in Chart 10.1 that an area in each tail of the distribution has been shaded. Each of these areas contains 2.5 percent or  $100(\alpha/2)$  percent, of the total

**CHART 10.1: DISTRIBUTION OF SAMPLE MEANS OF LENGTHS OF BOLT BLANKS TAKEN AT RANDOM FROM A LARGE POPULATION WITH  $\mu_0 = 200$  mm. AND  $\sigma = 20$  mm.**



area under the normal curve, and therefore their combined area is 5 percent, or  $100\alpha$  percent, of the total area under the normal curve. The test is said to be *two-sided* (or two-tailed), since we consider deviations in either direction from  $\mu_0$  to be important. Hence, we have shaded  $100(\alpha/2)$  percent of the area under the normal curve in each tail of the distribution to allow for both positive and negative deviations from  $\mu_0$ .

The values of  $\bar{X}$  and/or  $z$  which bound these two shaded areas are easily found. Using Appendix 3, we find a value  $z_U$  such that  $Q(z) = \alpha/2 = 0.025$ . We will call this the *upper rejection limit* for  $z$  and denote it  $z_U$ , where

$$z_U = z_{\alpha/2} = z_{0.025} = 1.96$$

The *lower rejection limit* for  $z$  is a value of  $z$  such that  $P(z) = 0.025$ . We will denote this value  $z_L$ , and because of the symmetry of the normal curve

$$z_L = z_{1-\alpha/2} = -z_U = -1.96$$

The corresponding upper rejection limit for  $\bar{X}$  is

$$\bar{X}_U = \mu_0 + z_U \sigma_{\bar{X}}$$

or

$$\bar{X}_U = 200 + 1.96(2) = 203.92$$

Similarly, the lower rejection limit for  $\bar{X}$  is

$$\bar{X}_L = \mu_0 + z_L \sigma_{\bar{X}}$$

or

$$\bar{X}_L = 200 - 1.96(2) = 196.08$$

since  $z_L$  is negative.

The shaded area in Chart 10.1 is called the *rejection region* (or critical region) of the test. The unshaded area is called the *acceptance region* of the test. It should be clear that the probability is  $\alpha = 0.05$  that an observed sample mean, drawn at random from the specified population, will be by chance as large as 203.92 or larger, or as small as 196.08 or smaller. Since the observed sample mean,  $\bar{X} = 205$ , is larger than  $\bar{X}_U = 203.92$ , the probability is less than  $\alpha/2 = 0.025$  that a sample mean of  $\bar{X} = 205$  or larger would be drawn at random from the specified population. Therefore, we *reject* the null hypothesis and reject the lot. Thus the decision rule is, "Take a random sample of 100 items; accept  $H_0$  if  $196.08 < \bar{X} < 203.92$ ; otherwise, reject  $H_0$ ."

We may also say that since the observed value of  $z$  is 2.5

$$z = \frac{\bar{X} - \mu_0}{\sigma_{\bar{X}}} = \frac{205 - 200}{2} = 2.5$$

and  $Q(z) = 0.00621$ , the probability of observing a value of  $z = 2.5$  or larger is 0.00621, which is less than  $\alpha/2 = 0.025$ ; and the probability of obtaining a *deviation* from  $\mu_0$  as large or larger than 5 mm is  $2(0.00621) = 0.01242$ , which is less than  $\alpha$ . Using either of these lines of reasoning, we reject  $H_0$ .

It should be remembered that in testing our hypothesis, two assumptions were made.

1. The sample was selected by simple random sampling. If the sampling were not random, a large observed deviation of  $\bar{X}$  from  $\mu_0$  would not be a basis for rejecting the null hypothesis.
2. The distribution of sample means is normal. The central limit theorem, however, tells us that even though the population is not normal, the distribution of sample means will be almost normal unless the sample size is rather small.

## 10.2 ONE-SIDED TEST

In most business situations we are not usually interested in testing whether a set of data comes from a population with a specified mean.

Indeed, it is incredible that a population mean could be exactly that specified.<sup>(1)</sup> We are usually interested in testing whether the population mean is *at least as large* as some specified value (or perhaps *at least as small* as some specified value). Consider the following decision problem.

An automobile tire manufacturer is considering changing the cord used in his tires. He wishes to change cord only if it can be demonstrated beyond reasonable doubt that the new cord (called Supertwist) has a flex life greater than 135 minutes. Thus, the hypotheses are

$$H_0: \mu = 135 \text{ min}$$

$$H_1: \mu > 135 \text{ min}$$

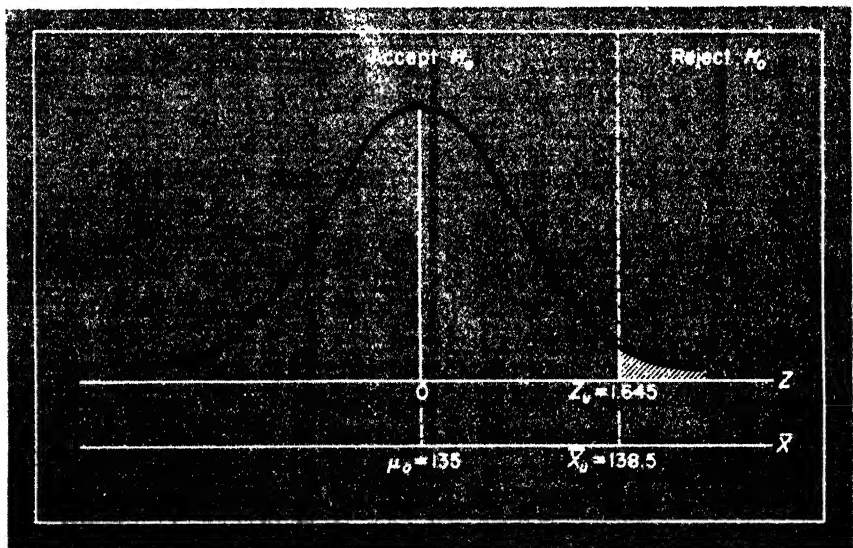
and we will refer to 135 min as  $\mu_0$ . The population standard deviation is known to be 15 min, and the manufacturer wishes to reject  $H_0$  only if the sample mean differs *positively* from  $\mu_0$  to such an extent that the probability of a difference as great as that observed or greater is less than or equal to 0.05. Thus, as before, we set  $\alpha = 0.05$  (where  $\alpha$  is the probability of a type I error).

A random sample of  $n = 50$  is taken from a very large lot of cord and found to have a mean flex life of

$$\bar{X} = 138.64 \text{ min}$$

Chart 10.2 shows the distribution of sample means, based upon a sample size of  $n = 50$ , taken from a very large lot of cord with  $\sigma = 15$  min and

**CHART 10.2: DISTRIBUTION OF SAMPLE MEANS OF FLEX LIFE OF TIRE CORD TAKEN AT RANDOM FROM A LARGE POPULATION WITH  $\mu_0 = 135$  min. AND  $\sigma = 15$  min.**



<sup>(1)</sup> Technically, if the variable is continuous the probability that  $\mu = \mu_0$  is zero.

$\mu_0 = 135$  min. Using Appendix 3, we find that 5 percent of the area under the right tail of the standardized normal curve is bounded from the left by  $z_U = z_\alpha = z_{0.05} = 1.645$ . In terms of the sample mean

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{50}} = 2.12$$

and the *single* value of  $\bar{X}$  that bounds the rejection region is

$$\bar{X}_U = \mu_0 + z_U \sigma_{\bar{X}}$$

or

$$\bar{X}_U = 135 + 1.645(2.12) = 138.5$$

Since the observed sample mean is  $\bar{X} = 138.64$ , we see that it is located in the rejection region, and we reject the null hypothesis. Alternatively, we may compute the observed value of  $z$

$$z = \frac{\bar{X} - \mu_0}{\sigma_{\bar{X}}}$$

or

$$z = \frac{138.64 - 135}{2.12} = 1.72$$

and reject  $H_0$  since  $z > z_U$ ; i.e.,  $1.72 > 1.645$ .

To recapitulate, we reject the null hypothesis because the probability of a chance occurrence of an  $\bar{X}$  or  $z$  value that is as large or larger than the one observed, if  $\mu = \mu_0$ , is less than alpha. Notice carefully that the reasoning used in this section is very similar to that used in the previous one. The principal difference is that the rejection region is all included in one tail of the distribution since we are, in this example, interested only in a positive difference from  $\mu_0$ . Thus, the test is said to be one-sided. If we were interested in a negative deviation from  $\mu_0$ , the rejection region would have been defined in the left tail of the distribution (see Problem 1).

### 10.3 THE POWER OF A TEST CONCERNING $\mu$

It should be recalled that  $\alpha$ , the probability of a type I error, is the probability of rejecting  $H_0$  when  $H_0$  is true. Thus, as we have defined it, alpha is the probability of rejecting  $H_0$  when  $\mu = \mu_0$ . It is obvious that a single number may be assigned to alpha.

If we define  $\beta(\mu)$  as the probability of accepting  $H_0$ , it is obvious that a single number cannot be assigned to  $\beta(\mu)$ , since  $\beta$  is a function of  $\mu$ . By definition,  $\beta(\mu_0)$  is the probability of accepting  $H_0$ , given that  $\mu = \mu_0$ . Hence,  $\beta(\mu_0) = 1 - \alpha$  and  $1 - \beta(\mu_0) = \alpha$ . When  $H_0$  is false,  $\beta(\mu)$  is the probability of a type II error.

The function  $1 - \beta(\mu)$  is called the *power function* for a test concerning the arithmetic mean. This function indicates the probability of rejecting  $H_0$  for

each possible value of  $\mu$ , given that  $\alpha$  and the rejection region are fixed. A power curve is the graph of the power function.<sup>(2)</sup>

**Two-sided Test.** Referring to our example dealing with bolt blanks, we recall that the rejection region was divided into two equal parts: an upper part bounded on the left by  $z_U = 1.96$  and a lower part bounded on the right by  $z_L = -1.96$ . Now if  $H_0$  is true, i.e., if  $\mu = 200$  mm, we know that we will nevertheless reject  $H_0$  5 percent of the time, since 5 percent of the sample means will be located in the rejection region by chance *even though*  $\mu = 200$  mm. Thus we know that

$$\alpha = P(z_L) + Q(z_U) = 0.025 + 0.025 = 0.05$$

Suppose now that  $\mu = 202$  mm. Clearly the null hypothesis is false. What is the probability of rejecting  $H_0$ , given that  $H_0$  is false; i.e., what is  $1 - \beta(\mu)$ ? We know that if an observed value of  $\bar{X}$  is as great or greater than  $\bar{X}_U = 203.92$  or as small or smaller than  $\bar{X}_L = 196.08$ , we will reject  $H_0$ . If  $\mu = 202$  mm and  $\sigma_{\bar{X}} = 2$ , the standardized difference between  $\bar{X}_L$  and  $\mu = 202$  mm is

$$z_1 = \frac{\bar{X}_L - \mu}{\sigma_{\bar{X}}} = \frac{196.08 - 202}{2} = -2.96$$

The standardized difference between  $\bar{X}_U$  and  $\mu = 202$  mm is

$$z_2 = \frac{\bar{X}_U - \mu}{\sigma_{\bar{X}}} = \frac{203.92 - 202}{2} = 0.96$$

When  $\mu = \mu_0$ ,  $z_1 = z_L$  and  $z_2 = z_U$ .

The probability that an observed value of  $\bar{X}$  will be located in the rejection region when  $\mu = 202$  mm is

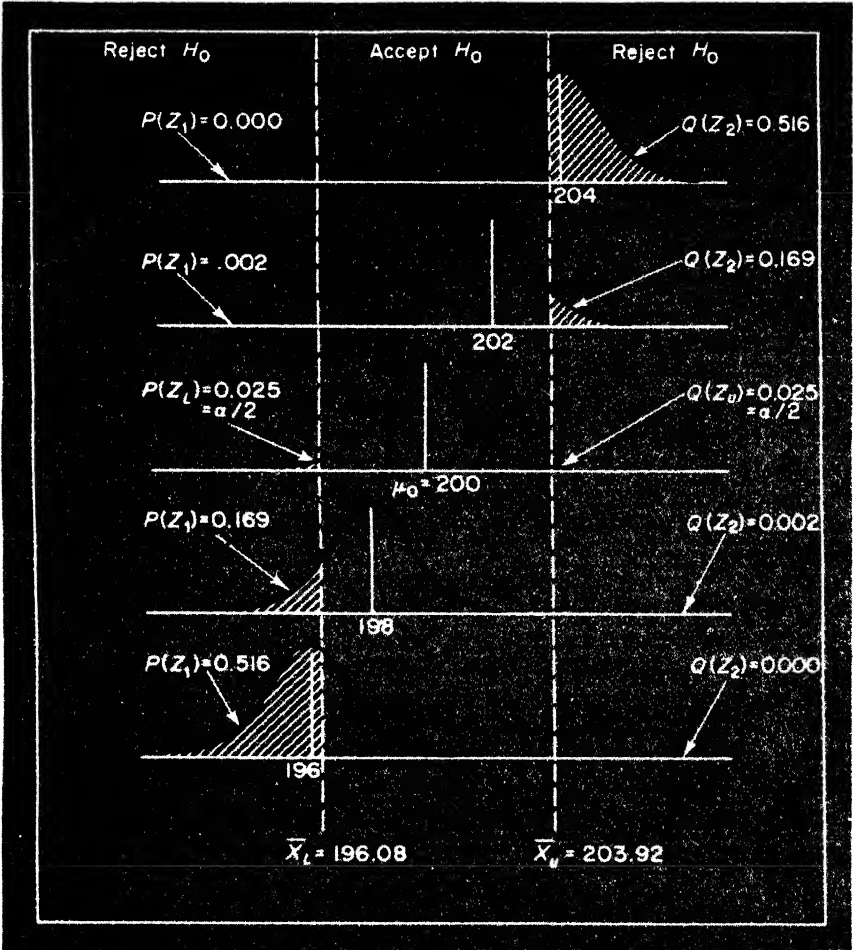
$$\begin{aligned} 1 - \beta(\mu) &= P(z_1) + Q(z_2) \\ \text{or } 1 - \beta(202) &= P(-2.96) + Q(0.96) \\ &= 0.002 + 0.169 = 0.171 \end{aligned}$$

Chart 10.3 illustrates how these probabilities vary with different selected values of  $\mu$ . A tabular form of calculation is given in Table 10.1 and the power function itself is graphed in Chart 10.4. The power curve in Chart 10.4 is based upon the calculations given in Table 10.1. Although there are infinitely many points, the curve is based only upon those points actually calculated. Notice that

$$1 - \beta(\mu_0) = \alpha$$

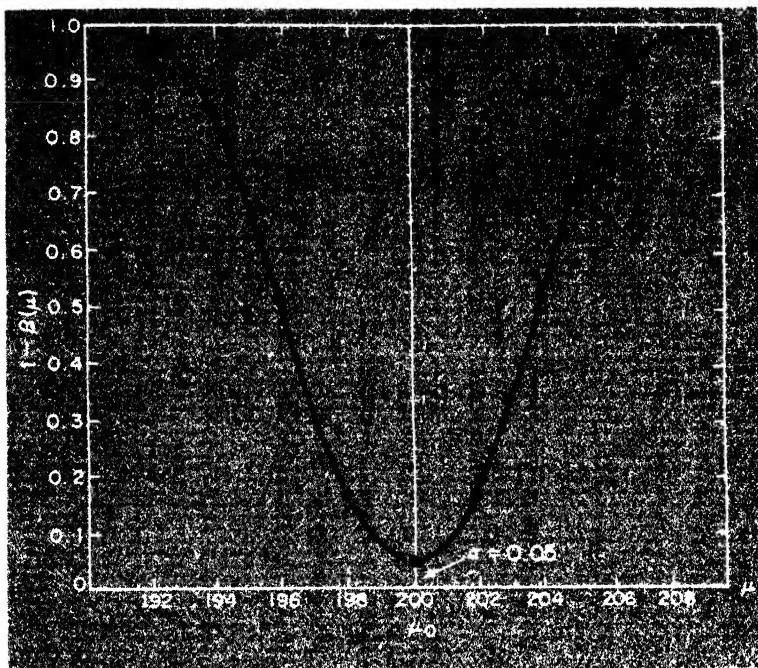
<sup>(2)</sup> Instead of the power function, some prefer to use the operating characteristic function of the test, O.C. The O.C. function is the complement of the power function; i.e., it is  $\beta(\mu)$ .

**CHART 10.3: PROBABILITY OF REJECTING  $H_0$  FOR SPECIFIED VALUES OF  $\mu$ , TWO-SIDED TEST.**



**TABLE 10.1: COMPUTATION OF SELECTED VALUES OF A POWER FUNCTION FOR A TWO-SIDED TEST CONCERNING  $\mu$**   
 $(\bar{X}_L = 196.08; \bar{X}_U = 203.92; \sigma_{\bar{X}} = 2)$

$\mu$	$z_1 = (\bar{X}_L - \mu)/\sigma_{\bar{X}}$	$P(z_1)$	$z_2 = (\bar{X}_U - \mu)/\sigma_{\bar{X}}$	$Q(z_2)$	$1 - \beta(\mu) = P(z_1) + Q(z_2)$
192	2.04	0.979	5.96	0.000	0.979
194	1.04	0.851	4.96	0.000	0.851
196	0.04	0.516	3.96	0.000	0.516
198	-0.96	0.169	2.96	0.002	0.171
200 = $\mu_0$	-1.96	0.025	1.96	0.025	0.050 = $\alpha$
202	-2.96	0.002	0.96	0.169	0.171
204	-3.96	0.000	-0.04	0.516	0.516
206	-4.96	0.000	-1.04	0.851	0.851
208	-5.96	0.000	-2.04	0.979	0.979

**CHART 10.4: POWER OF A TWO-SIDED TEST CONCERNING THE ARITHMETIC MEAN.**

**One-sided Test.** In the tire cord example we said that the rejection region was bounded on the left by  $\bar{X}_U = 138.5$  min or  $z_U = z_\alpha = 1.645$ . If the null hypothesis is true, i.e., if  $\mu = 135$  min,

$$\alpha = Q(z_U) = 0.05$$

Suppose, however, that  $\mu = 140$  min. Then  $\bar{X}_U$  is  $-0.71$  standard errors below  $\mu = 140$  min. Recalling that  $\sigma_{\bar{X}} = 2.12$ , we have

$$z_2 = \frac{\bar{X}_U - \mu}{\sigma_{\bar{X}}} = \frac{138.5 - 140}{2.12} = -0.71$$

and

$$1 - \beta(\mu) = Q(z_2)$$

or

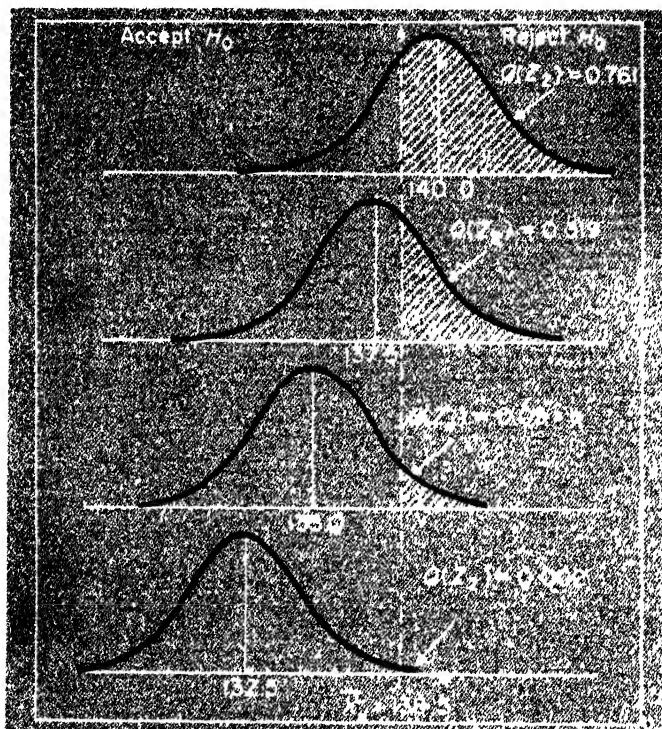
$$1 - \beta(140) = Q(-0.71) = 1 - Q(0.71) = 0.761$$

Chart 10.5 illustrates how these probabilities vary with different selected values of  $\mu$ . A tabular form of calculation is given in Table 10.2, and the power function itself is graphed in Chart 10.6. The power curve in Chart 10.6 is based upon the calculations given in Table 10.2. Notice that

$$1 - \beta(\mu_0) = \alpha$$



**CHART 10.5: PROBABILITY OF REJECTION OF  $H_0$  FOR SPECIFIED VALUES OF  $\mu$ , ONE-SIDED TEST.**

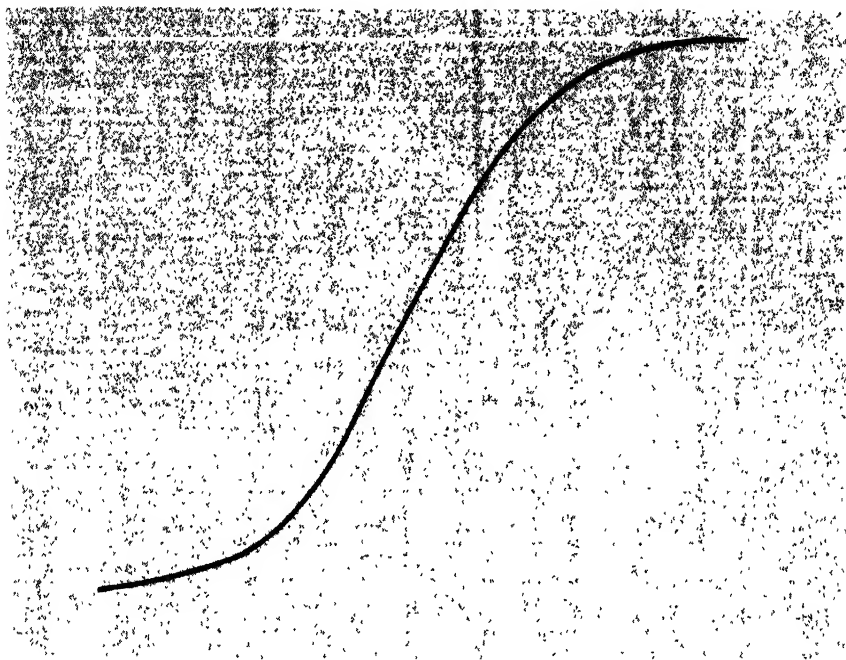


**TABLE 10.2: COMPUTATION OF SELECTED VALUES OF A POWER FUNCTION FOR A ONE-SIDED TEST CONCERNING  $\mu$**

( $\bar{X}_0 = 138.5$ ;  $\sigma_{\bar{X}} = 2.12$ )

$\mu$	$z_1 = \frac{(\bar{X}_0 - \mu)/\sigma_{\bar{X}}}{1}$	$1 - \beta(\mu) = Q(z_1)$
132.5	2.83	0.002
135.0 = $\mu_0$	1.65	0.050 = $\alpha$
137.5	0.47	0.319
140.0	-0.71	0.761
142.5	-1.89	0.971
145.0	-3.07	0.999

In other words, if we had perfect information concerning the value of  $\mu$  we would always reject the null hypothesis when the alternative hypothesis is true and never reject it otherwise. However, because we cannot have perfect

**CHART 10.6: POWER OF A ONE-SIDED TEST CONCERNING THE ARITHMETIC MEAN.**

**Comments on the Power of a Test.** For a test of the hypotheses

$$H_0: \mu = \mu_0$$

$$H_1: \mu > \mu_0$$

we have seen that  $1 - \beta(\mu)$  increases as  $\mu$  departs positively from  $\mu_0$ . An “ideal” power function for such a test would be<sup>(3)</sup>

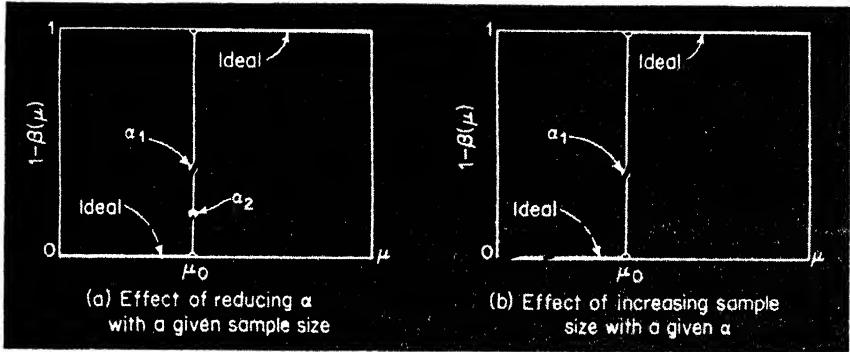
$$1 - \beta(\mu) = 0, \quad \text{when } \mu < \mu_0$$

$$1 - \beta(\mu) = 1, \quad \text{when } \mu > \mu_0$$

information about  $\mu$  without enumerating the entire population, this ideal power function cannot be attained. Furthermore, with a fixed sample size, we cannot approach the ideal more closely on one side of  $\mu_0$  without moving away from the ideal on the other side. Chart 10.7 illustrates this point. The figure on the left shows that reducing  $\alpha$  (moving  $\bar{X}_U$  further to the right of

<sup>(3)</sup> We are willing to accept the null hypothesis,  $H_0: \mu = \mu_0$ , even though  $\mu < \mu_0$ .

CHART 10.7: COMPARISONS OF POWER CURVES FOR ONE-SIDED TEST.



$\mu_0$ ) will move the power curve downward, thus giving the dotted power curve. This power curve is closer to the ideal to the left of  $\mu_0$  but further away from the ideal to the right of  $\mu_0$ . In other words, by reducing the  $\alpha$  we have reduced  $1 - \beta(\mu)$  everywhere. The figure on the right in Chart 10.7 shows the behavior of the power curve as the sample size is increased. The power curve becomes steeper and more closely approaches the ideal on both sides of  $\mu_0$ .

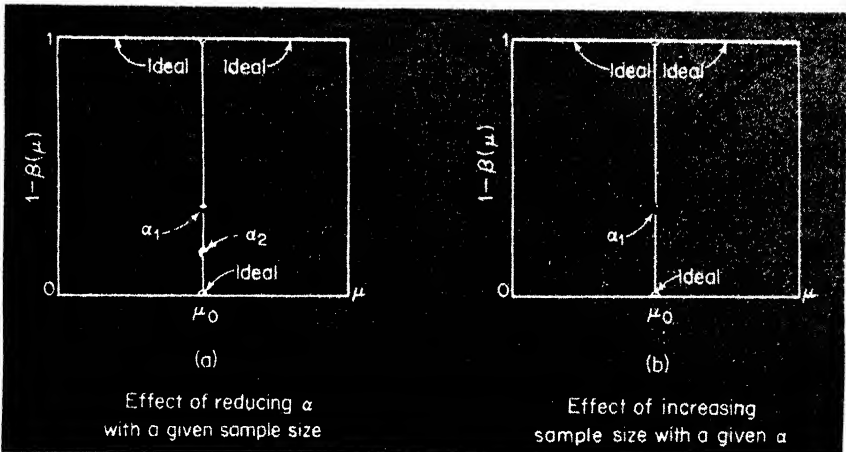
Similar remarks can be made about the two-sided test. Here, the ideal power function is

$$1 - \beta(\mu) = 0, \text{ when } \mu = \mu_0$$

$$1 - \beta(\mu) = 1, \text{ when } \mu \neq \mu_0$$

The figure to the left in Chart 10.8 shows that  $1 - \beta(\mu)$  is shifted downward everywhere when  $\alpha$  is reduced, given a fixed sample size. The figure on the right in Chart 10.8 shows that  $1 - \beta(\mu)$  can be brought closer to the ideal by

CHART 10.8: COMPARISONS OF POWER CURVES FOR TWO-SIDED TEST.



increasing the sample size. As the sample size approaches the population size, the power curve collapses on the ideal. In both cases the ideal power curves are approached with an increase in the sample size, because as  $n$  approaches  $N$  the standard error of the mean approaches zero. Tests that allow control of both  $\alpha$  and  $\beta(\mu)$  are considered in an appendix to this chapter.

## 10.4 CONFIDENCE LIMITS

As was explained in Chapter 8, in addition to making a point estimate of  $\mu$  on the basis of a single sample mean, we often make an interval estimate of  $\mu$  and attach some degree of confidence to this estimate. The measure of the confidence that we have in our interval estimate is called a *confidence coefficient* and is often stated  $1 - \alpha$ . To determine the  $100(1 - \alpha)$  percent *confidence limits* of  $\mu$  with  $\sigma$  known, we find:

1. A value of  $\mu_1$  less than  $\bar{X}$ , such that  $\bar{X}$  cuts off the *upper*  $(100 \alpha/2)$  percent of the tail of the normal distribution associated with  $\mu_1$ .
2. A value of  $\mu_2$  greater than  $\bar{X}$ , such that  $\bar{X}$  cuts off the *lower*  $(100 \alpha/2)$  percent of the tail of the normal distribution associated with  $\mu_2$ .

Thus, if we wish 95 percent confidence limits,  $1 - \alpha = 0.95$ ,  $\alpha = 0.05$ , and  $\mu_1$  is so determined that  $\bar{X}$  cuts off the upper 2.5 percent of the tail of the distribution associated with  $\mu_1$ , while  $\mu_2$  is so determined that  $\bar{X}$  cuts off the lower 2.5 percent of the tail of the distribution associated with  $\mu_2$ .

The two values of  $\mu$ , the lower and upper confidence limits, are both obtained from the expression

$$\bar{X} - \mu_1 = \mu_2 - \bar{X} = z_{\alpha/2} \sigma_{\bar{X}} \quad (10-1)$$

The 95 percent confidence limits for the tire cord data are easily obtained.

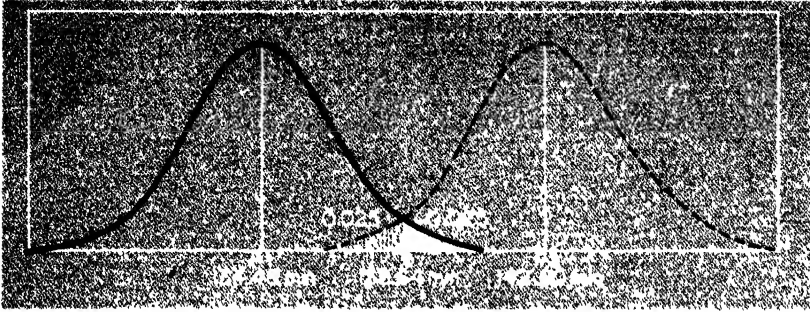
$$\begin{aligned} 138.64 - \mu_1 &= 1.96(2.12) = 4.16 \\ \mu_1 &= 138.64 - 4.16 = 134.48 \text{ min} \\ \mu_2 - 138.64 &= 1.96(2.12) = 4.16 \\ \mu_2 &= 138.64 + 4.16 = 142.80 \text{ min} \end{aligned}$$

Thus the 95 percent confidence limits are 134.48 min and 142.80 min and the *confidence interval* is

$$\mu_2 - \mu_1 = 142.8 \text{ min} - 134.48 \text{ min} = 8.32 \text{ min}$$

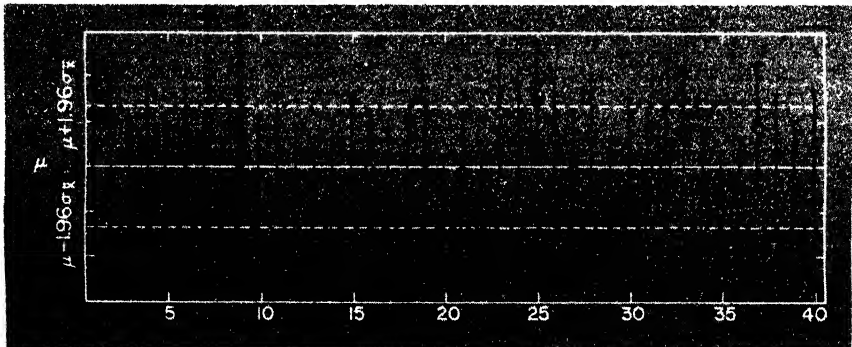
This procedure is a little easier to understand after examination of Chart 10.9. If  $\mu_1 = 134.48$  min, the probability that  $\bar{X}$  will be as large as 138.64 min or larger is  $\alpha/2 = 0.025$ . If  $\mu_2 = 142.80$  min, the probability that  $\bar{X}$  will be as small as 138.64 min or smaller is  $\alpha/2 = 0.025$ .

**CHART 10.9: UPPER AND LOWER 95 PER CENT CONFIDENCE LIMITS FOR  $\mu$ :  $n = 50$ ;  $\bar{X} = 138.64$  min.;  $\sigma = 15$  min.**



We have just made an estimate of the 95 percent confidence limits of  $\mu$  from *one sample*. The true but unknown value of  $\mu$  may or may not lie within the limits of 134.48 min and 142.80 min and it is illogical (using the classical definition of probability) to say that the probability is 0.95 that  $\mu$  is within these limits, since  $\mu$  either lies within the limits or does not. But if we made many such determinations of the 95 percent confidence limits of  $\mu$ , each from a *different sample*, we would find that our many sets of limits would include  $\mu$  about 95 times in 100 and fail to include  $\mu$  about 5 times out of 100. The reason that this is true is that the means of 95 out of 100 samples are within  $\pm 1.96\sigma_{\bar{X}}$  of  $\mu$ , and the confidence limits are determined by measuring  $\pm 1.96\sigma_{\bar{X}}$  from the different sample means. Consider Chart 10.10. The solid line represents  $\mu$ . The dotted lines are  $1.96\sigma_{\bar{X}}$  from  $\mu$ . Of the 40 samples whose means are shown by circles, 38 (or 95 percent) of the sample means are inside the dotted lines; two, shown in black, are outside the dotted lines. For each sample,  $\mu_1$  and  $\mu_2$  are shown by horizontal bars. These horizontal bars are at distances of  $1.96\sigma_{\bar{X}}$  from  $\bar{X}$ . Each vertical line of Chart 10.10 corresponds to the base line of Chart 10.9. Only in the case of the two black sample means is the line representing  $\mu$  not enclosed.

**CHART 10.10: 95 PER CENT CONFIDENCE LIMITS FOR MEAN, AND 40 SAMPLES, STANDARD DEVIATION KNOWN.**



A correct probability statement concerning 95 percent confidence limits is: "Before we select a sample, the probability is 0.95 that the confidence limits we compute from the sample will enclose the population mean."

If the different confidence limits are desired, the procedure would be identical, but different  $z$  values would be used. For 99 percent limits,  $\alpha/2 = 0.005$ , and  $z_{\alpha/2} = 2.576$ .

The confidence interval varies directly with the standard error of the mean. Therefore, the confidence interval varies directly with the standard deviation of the population and inversely with the square root of the sample size.

Stating confidence limits may be thought of as a substitute for testing a hypothesis. If  $\mu_0$  is between the confidence limits,  $H_0$  is accepted; if  $\mu_0$  is on or outside the confidence limits,  $H_0$  is rejected.<sup>(4)</sup> Confidence limits of stated width are considered in an appendix to this chapter.

## 10.5 ON SETTING ALPHA

The value of alpha must be set in advance of testing a hypothesis or setting confidence limits. The smallest value alpha may take is zero, in which case the null hypothesis would never be rejected. The largest value alpha may take is one, in which case the null hypothesis would always be rejected. If alpha is one, the confidence limits will both be  $\bar{X}$  (the confidence coefficient will be zero) and if alpha is zero, confidence limits will be  $\pm \infty$  (the confidence coefficient will be one). This discussion, of course, assumes imperfect knowledge of  $\mu$ .

Traditionally, alpha is usually set at a level such as 0.005, 0.01, 0.025, 0.05, or 0.10, and a null hypothesis that is rejected at  $\alpha = 0.01$  or smaller is sometimes said to be "highly significant."

Traditional statistical inference does not offer any firm rules for the setting of alpha. Rather, it is pointed out that two considerations are involved.

1. The greater the degree of belief in the null hypothesis, the smaller should alpha be set.
2. The greater the cost of rejecting the null hypothesis, given that it is true, the smaller should alpha be set.

Both of these considerations are complex. In general, all that is said is that we wish to cause the probability of a type I error to be rather small for well-established null hypotheses and for null hypotheses whose erroneous rejection would be very costly. We will consider these matters in greater detail in Chapter 13, where there is explicit recognition of costs and prior degrees of belief in the null hypothesis.

---

<sup>(4)</sup> This statement has reference to a two-sided test of hypotheses and a central confidence limit. If the test of the hypothesis is one-sided, the confidence limit is also one-sided;  $\mu_1$  or  $\mu_2$  is stated, but not both.

## PROBLEMS

1. Given that

$$\sigma = 10, \quad n = 64, \quad \bar{X} = 43, \quad \alpha = 0.10$$

test

$$a. H_0: \mu = 45 \quad b. H_0: \mu = 45$$

$$H_1: \mu \neq 45 \quad H_1: \mu < 45$$

Evaluate 5 points on the power curves associated with  $a$  and  $b$  and draw these power curves.

2. Using the information in Problem 1, set 90 percent confidence limits for  $\mu$ . Explain what these limits mean. If  $\sigma = 20$ , other things the same, would the limits be different? Why? What effect does the sample size have on your limits?

3. Suppose you were testing the family

$$H_0: \mu = \mu_0$$

$$H_1: \mu < \mu_0$$

and set your rejection region in the right tail of the distribution of sample means. How would your power curve appear? Generalize your results.

4. Using the information in Problem 1, indicate the way in which your power curves would change:

a. If  $n = 144$ .

b. If  $\alpha = 0.05$ .

c. If  $n = \infty$ .

d. If  $\alpha = 0$ .

Explain in words the reasons for these changes.

## APPENDIX: Determination of Sample Size Risk through Control

The methods of Chapter 10 assume a predetermined sample size. Under these conditions it was shown that both  $\alpha$  and  $\beta(\mu)$  could not be controlled at the same time. We now consider control of both  $\alpha$  and  $\beta(\mu)$  through the selection of an appropriate sample size.<sup>(1)</sup>

<sup>(1)</sup> We continue to assume that the sample size will be large enough to insure that the distribution of sample means will be approximately normal.

### A10.1 ONE-SIDED TEST CONTROLLING BOTH $\alpha$ AND $\beta(\mu)$

Suppose for our tire cord example of Section 10.2 that, as before, the manufacturer wishes to reject the null hypotheses

$$H_0: \mu = 135 \text{ min}$$

in favor of the alternative

$$H_1: \mu > 135 \text{ min}$$

only if the observed sample mean differs positively from 135 min to such an extent that the probability of a difference as large or larger than that observed is less than or equal to 0.05. Thus at

$$\mu_0 = 135 \text{ min}, \quad \alpha = 0.05$$

In addition to this protection the manufacturer also wants assurance that *when* the population mean is 140 min he will accept the null hypothesis (which of course will be false) at most 10 times in 100. Thus at

$$\mu_1 = 140 \text{ min} \quad \beta(\mu_1) < 0.10$$

The lower part of Chart A10.1 shows that the desired protection against both types of error may be thought of as specifying two points on a power curve. The upper part of this chart shows the two sampling distributions associated with  $\mu_0$  and  $\mu_1$ . The dotted vertical line shows the critical value of  $\bar{X}$ , as yet unknown, which bounds from below 5 percent of the right tail of the distribution associated with  $\mu_0 = 135$  min and *at the same time* bounds from above 10 percent of the area in the left tail of the distribution associated with  $\mu_1 = 140$  min. Thus for the desired risk control we need to find the minimum value of  $n$  and an appropriate value for  $\bar{X}_c$ .

Referring to Chart A10.1, we see that

$$\bar{X}_c = \mu_0 + z_\alpha \sigma_{\bar{X}}$$

and

$$\bar{X}_c = \mu_1 + z_\beta \sigma_{\bar{X}}$$

where  $z_\alpha$  and  $z_\beta$ , respectively, cut off the upper 5 percent of the distribution associated with  $\mu_0$  and the lower 10 percent of the distribution associated with  $\mu_1$ . Of course,  $z_\beta$  will be negative.

From our knowledge that

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

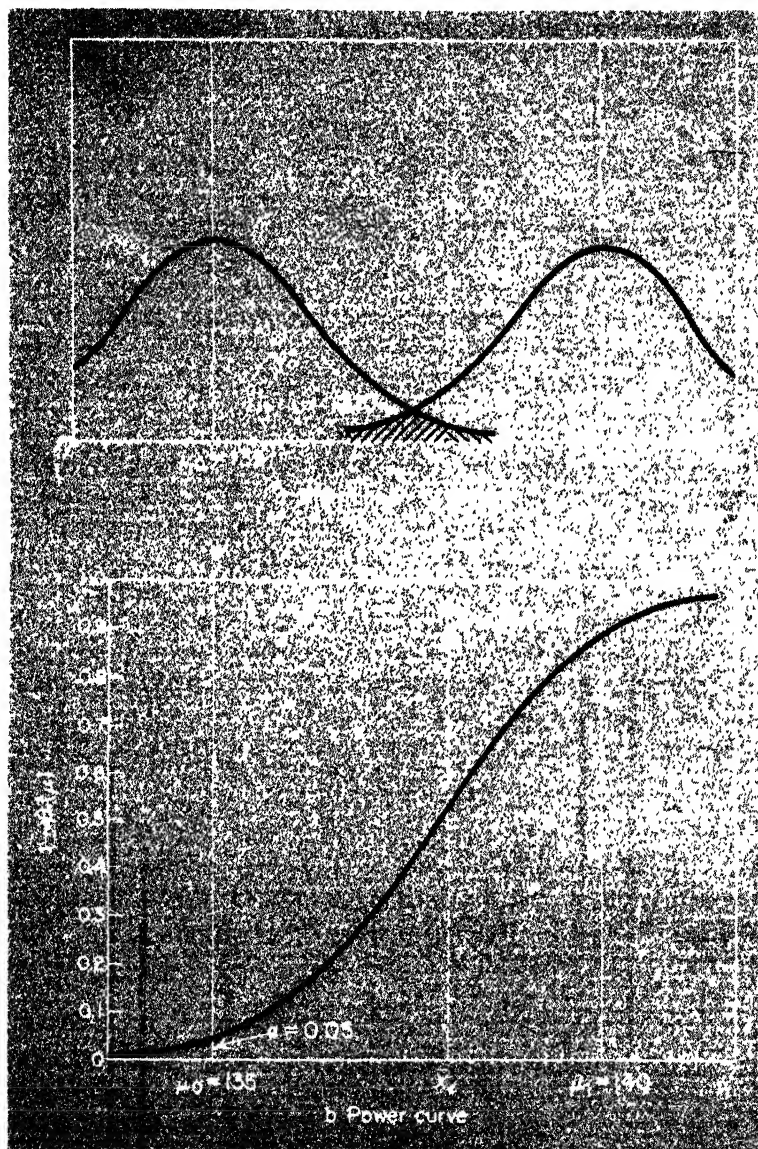
we can write the above two equations as

$$\bar{X}_c = \mu_0 + z_\alpha \left( \frac{\sigma}{\sqrt{n}} \right) \quad (\text{A10-1})$$

$$\bar{X}_c = \mu_1 + z_\beta \left( \frac{\sigma}{\sqrt{n}} \right) \quad (\text{A10-2})$$



**CHART A10.1: DISTRIBUTION OF MEANS AND POWER CURVE ASSOCIATED WITH A ONE-SIDED TEST WHERE:  $\mu_0 = 135$ ;  $\alpha = 0.05$ ;  $\mu_1 = 140$ ;  $\beta(\mu_1) = 0.10$ .**



and if we set these two equations equal to each other and solve for  $n$ , we find that

$$n = \left[ \frac{(z_\alpha - z_\beta)\sigma}{\mu_1 - \mu_0} \right]^2$$

Entering Appendix 3, we find

$$z_\alpha = z_{0.05} = 1.645 \quad \text{and} \quad z_\beta = -z_{0.10} = -1.282$$

so that

$$n = \left[ \frac{(1.645 + 1.282)15}{140 - 135} \right]^2 = 77.1$$

since  $\sigma = 15$  minutes by the specification in Sec. 10.2. Then, using either Eqs. (A10-1) or (A10-2), we may solve for  $\bar{X}_c$ . For example, from Eq. (A10-1) we find

$$\bar{X}_c = 135 + 1.645 \frac{15}{\sqrt{77.1}} = 137.8 \text{ min}$$

One problem with this analysis is that the sample size will generally not turn out to be an integer. In most cases it will be best to round the calculated sample size to the next *largest* integer so that both  $\alpha$  and  $\beta(\mu_1)$  will be at least as small as desired. The student may verify by calculation of the two points on the power curve that both  $\alpha$  and  $\beta(\mu_1)$  are smaller than required when  $n = 78$ . Of course, a recalculation of  $\sigma_{\bar{X}}$  is necessary after  $n$  is rounded upward. If we wish  $\alpha = 0.05$  and  $\beta(\mu_1) \leq 0.10$ , we must recalculate  $\bar{X}_c$  by substituting  $n = 78$  into Eq. (A10-1). The student may also wish to verify that the same type of analysis can be used in conducting a two-sided test.

## A10.2 DETERMINATION OF SAMPLE SIZE WITH CONFIDENCE LIMITS

Suppose that, using the example of Sec. 10.1, we wished to estimate the mean of the lot of bolt blanks within  $\pm 6$  mm, and (to keep the calculation simple) with a confidence coefficient of 99.73 percent. We know that

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

and that

$$z_{\alpha/2} = \frac{\bar{X} - \mu_0}{\sigma_{\bar{X}}}$$

where the confidence coefficient is  $100(1 - \alpha)$  percent. It follows, then, that we may combine these two equations and write

$$n = \left( \frac{z_{\alpha/2}\sigma}{\bar{X} - \mu_0} \right)^2$$

In the case of our example,  $\alpha = 0.0027$  and  $\alpha/2 = 0.00135$ . Entering Appendix 1, we find that  $z_{0.00135} = 3$  so that

$$n = \left[ \frac{3(20)}{6} \right]^2 = 100$$

since  $\bar{X} - \mu_0 = 6$  mm and  $\sigma = 20$  mm as specified. Again,  $n$  will usually not be an integer and should be rounded to the next larger integer to make  $(1 - \alpha)$  slightly *larger* than required.

# 11

## Tests of Hypotheses and Confidence Limits for the Arithmetic Mean: Population Variance Unspecified

In the preceding chapter we were concerned with some statistical procedures related to the arithmetic mean of a single sample when the population variance was known or specified. In this chapter we extend our discussion to the case where the population variance is not specified.<sup>(1)</sup> We will also consider tests involving two samples.

### 11.1 THE *t*-DISTRIBUTION

We have frequently made use of the formula for the variance of the mean

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n} \quad (11-1)$$

which enables us to determine the standard error of the mean for infinite or very large populations when the population variance is

---

<sup>(1)</sup> The methods of this chapter closely parallel those of the last chapter. However, the power function of the “*t*-test” is beyond the scope of this text. The power function of the *t*-test is explained and illustrated for selected values of alpha in: E. S. Pearson and H. O. Hartley (eds.) *Biometrika Tables for Statisticians*, Vol. I (Cambridge University Press, 3rd Ed.), pp. 24ff. Use of these functions allows risk control similar to that discussed in the appendix to Chapter 10.

known or specified. It should be noted, however, that most of the time  $\sigma^2$  is not known. In Sec. 4.4 and elsewhere we noted that the variance

$$s^2 = \frac{\sum x^2}{n - 1}$$

is an unbiased estimator of  $\sigma^2$ . It would seem reasonable to use  $s^2$  as an estimator of  $\sigma^2$  when  $\sigma^2$  is not known. Thus, to estimate  $\sigma_X^2$  we replace  $\sigma^2$  in Eq. (11-1) with  $s^2$  and form

$$s_X^2 = \frac{s^2}{n} \quad (11-2)$$

When  $\sigma^2$  is specified, we have used the statistic

$$z = \frac{\bar{X} - \mu}{\sigma_X}$$

This statistic is normally distributed if the population is normal or if the sample size is large. By analogy, when  $\sigma^2$  is estimated from the sample we form the statistic

$$t = \frac{\bar{X} - \mu}{s_X} \quad (11-3)$$

The question now becomes: What is the distribution of the “ $t$  ratio?”

The  $t$  distribution was formalized by W. S. Gosset, writing under the name “Student” in 1908, and the  $t$  ratio is often referred to as “Student’s  $t$ .” The  $t$  distribution is not a single distribution but a family of distributions, each member of which depends upon a quantity known as “the number of degrees of freedom.” In this text the number of degrees of freedom will be denoted by the Greek letter  $\nu$  (nu).

It is difficult to give a single definition of the term “degrees of freedom,” since this term has various shades of meaning in statistical literature. In general, it is probably best to define the number of degrees of freedom for a set of observations as the number of elements in the set that are free to vary, given the restrictions placed on the set by some statistic or statistics that have been computed from the set. For example, given a set of four observations, if the restriction is made in advance that the sum of the set is five, we know that one of the elements in the set is also specified, or restricted. Thus, if three of the four numbers are

1, 5, 8

we know that the fourth number *must* be  $-9$  for the sum of all four to be 5. In this example, the number of elements in the set that are free to vary is 3, which is the number of elements in the set,  $n$ , minus 1.

When  $s^2$  is used to estimate  $\sigma^2$ , the number of degrees of freedom is similarly  $n - 1$ . We reason this way because when we calculate  $s^2$  we must know the value of the arithmetic mean of the set of observations. Of the  $n$

items in the sample,  $n - 1$  can be selected arbitrarily. Once this selection has been made, the value of  $\bar{X}$  determines the remaining  $X$  value.

A  $t$  distribution is symmetrical but is more widely dispersed than the standardized normal distribution.<sup>(2)</sup> For the same value of  $z$  and  $t$  there is always as large as or larger a proportion of the area in the tail of the  $t$  distribution. Also, for the same proportion of area in the tail,  $t$  is larger than  $z$ . However, the differences between the standardized normal distribution and a given  $t$  distribution become less as the number of degrees of freedom upon which the given  $t$  distribution is based increases. When  $\nu = 100$  or larger, there is very little difference in the distributions, and the standardized normal distribution is usually substituted for the given  $t$  distribution. Many statisticians make the substitution when  $\nu > 30$ . The close relationship between the two distributions is clearly shown in Chart 11.1. When the number of degrees of freedom is small (shown in the upper part of the chart), the difference in the areas bounded by the same values of  $t$  and  $z$  is fairly large. When the number of degrees of freedom is rather large (shown in the lower part of the chart), the difference in the areas is fairly small. Because the difference between  $z$  and  $t$  is important only when  $\nu$  is small, methods employing the  $t$  distribution are often called small-sample methods. The theoretical distinction, however, is not whether the sample size (or  $\nu$ ) is small, but whether  $\sigma^2$  is given or estimated.

## 11.2 TWO-SIDED TEST

Continuing with the illustration begun in the last chapter concerning the lengths of bolt blanks, suppose that as before we wish to test

$$H_0: \mu = 200 \text{ mm}$$

$$H_1: \mu \neq 200 \text{ mm}$$

at  $\alpha = 0.05$ . Again,  $\mu_0 = 200$  mm. In this case suppose that we do not know  $\sigma^2$ . To estimate  $\sigma^2$  we will use  $s^2$  and recall that we may write

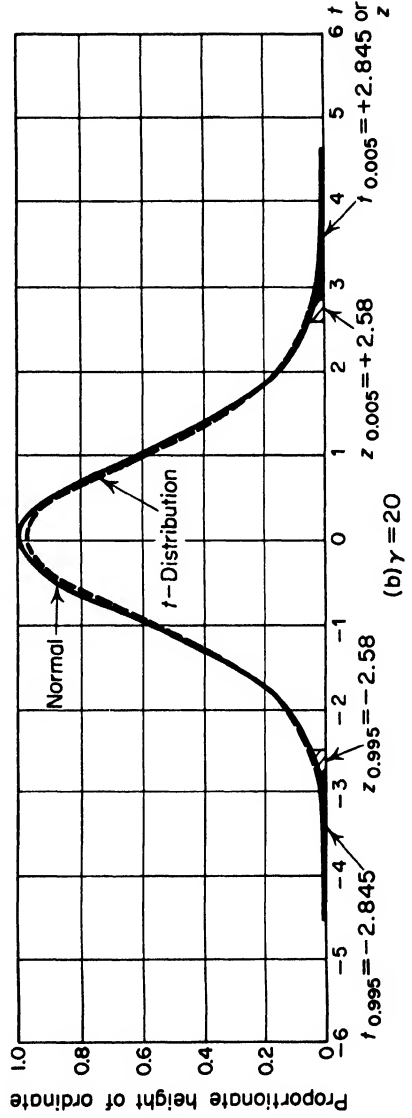
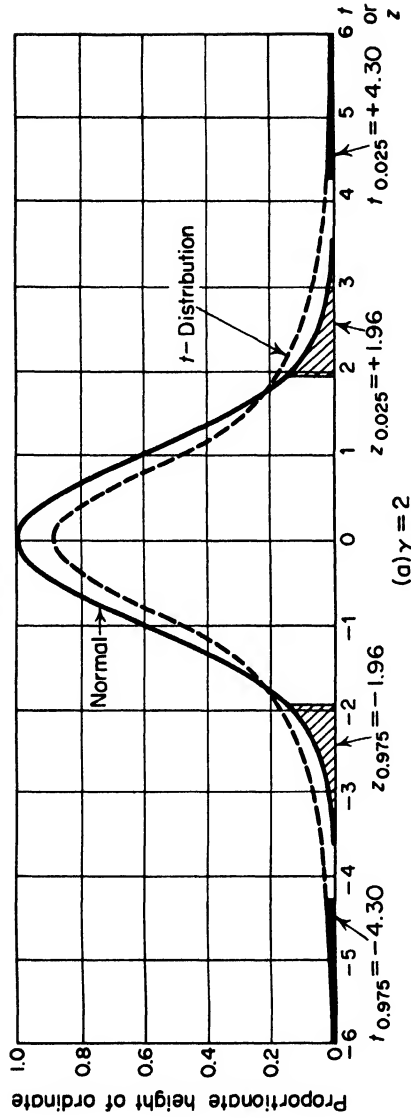
$$s^2 = \frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n - 1}$$

since  $\sum x^2 = \sum X^2 - (\sum X)^2/n$ .

---

<sup>(2)</sup> The variance of a  $t$  distribution is:  $\sigma^2 = \nu/(\nu - 2)$  for  $\nu > 2$ , and a measure of kurtosis (excess) is:  $\gamma_2 = 6/(\nu - 4)$  for  $\nu > 4$ . It is easy to see, therefore, that the variance of almost any  $t$  distribution is greater than the variance of the standardized normal distribution and that almost any  $t$  distribution is leptokurtic. In the limit, when  $\nu = \infty$ , the associated  $t$  distribution is identical to the standardized normal distribution with variance of one and no excess. A comparison of Appendix 1 (for the standardized normal distribution) and Appendix 4 (for the  $t$  distribution) will verify that  $z_q$  and  $t_q$  are the same when  $\nu = \infty$ .

CHART 11.1: COMPARISON OF *t* DISTRIBUTION WITH NORMAL DISTRIBUTION.



From a sample of  $n = 25$  items, drawn at random, suppose that we have calculated

$$\Sigma X = 5125 \text{ mm}$$

$$\Sigma X^2 = 1,058,401$$

Then 
$$s^2 = \frac{1,058,401 - \frac{(5125)^2}{25}}{24} = 324$$

and 
$$\bar{X} = \frac{5125}{25} = 205 \text{ mm}$$

Substituting the value of  $s^2$  into Eq. (11-2) gives

$$s_{\bar{X}}^2 = \frac{324}{25} = 12.96$$

and 
$$s_{\bar{X}} = \sqrt{12.96} = 3.6 \text{ mm}$$

Then, the observed value of  $t$  is

$$t = \frac{\bar{X} - \mu_0}{s_{\bar{X}}} = \frac{205 - 200}{3.6} = 1.39$$

and since  $n = 25$ ,  $\nu = n - 1 = 24$ .

Now we are dealing with  $t$  values rather than  $z$  values, and we refer to the “ $t$  table” given in Appendix 4. This table shows the number of degrees of freedom  $\nu$  in the first column, selected values of  $Q(t | \nu)$  across the top of the table, and values of  $t_Q$  in the body of the table. We stress again that the area in the right tail of a given  $t$  distribution, indicated symbolically by  $Q(t | \nu)$ , is a function of the number of degrees of freedom  $\nu$  upon which the given  $t$  distribution is based.

Using the same reasoning process as discussed in the last chapter, we seek a value of  $t_Q$  such that  $Q(t | \nu) = \alpha/2 = 0.025$ . From Appendix 4, with  $\nu = 24$ , we find the upper rejection value for  $t$ ,  $t_Q = t_{\alpha/2} = t_{0.025} = 2.064$ . Because of the symmetry of the  $t$  distribution,  $t_L = -t_U = -2.064$ .

Chart 11.2 shows the  $t$  distribution associated with this test. We see that the calculated value of  $t = 1.39$  does not fall into the rejection region and, therefore, we do *not* reject the null hypothesis. Alternatively, we may calculate the rejection values for  $\bar{X}$ :

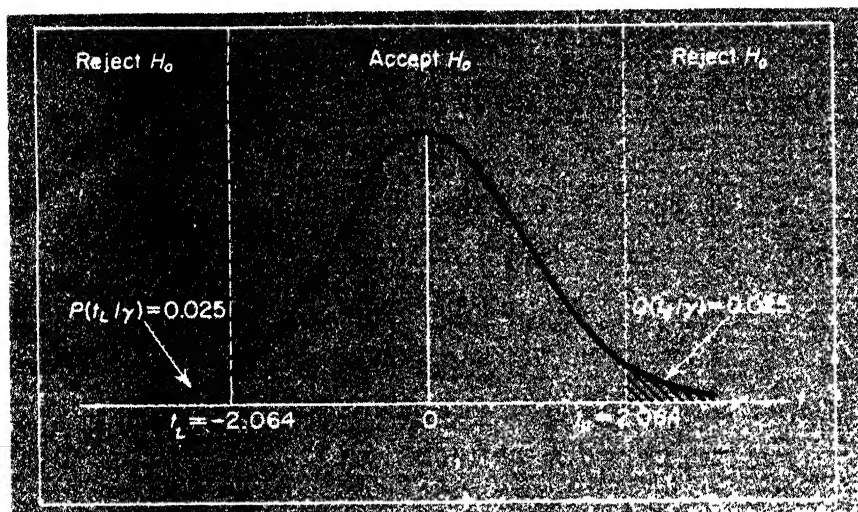
$$\begin{aligned}\bar{X}_L &= \mu_0 + t_L s_{\bar{X}} \\ &= 200 - 2.064(3.6) = 192.6 \text{ mm}\end{aligned}$$

and 
$$\begin{aligned}\bar{X}_U &= \mu_0 + t_U s_{\bar{X}} \\ &= 200 + 2.064(3.6) = 207.4 \text{ mm}\end{aligned}$$

Since  $\bar{X} = 205$  but  $\bar{X}_U = 207.4$ , we do not reject  $H_0$ . However, we must be careful *not* to say that the reason for this decision was that the deviation of  $\bar{X}$  from  $\mu_0$  was smaller than would have been expected by chance 95 percent of the time. It is correct, however, to say that a value of  $t$  was smaller than



**CHART 11.2: CRITICAL VALUES OF  $t$  FOR BOLT BLANK PROBLEM ( $\nu = 24$ ;) ( $s_{\bar{X}} = 3.6$ ;  $\alpha = 0.05$ ).**



would have been expected by chance 95 percent of the time. The  $t$  ratio is affected *not only* by the difference between  $\bar{X}$  and  $\mu_0$ , but also by the value of  $s_{\bar{X}}$ , which varies from sample to sample. When a value of  $t$  is associated with a small value of  $Q(t | \nu)$ , it may be that the deviation of  $\bar{X}$  from  $\mu_0$  is unusually large or that  $s_{\bar{X}}$  is unusually small, or both. Conversely, a large value of  $Q(t | \nu)$  may result from an unusually small value of  $\bar{X}$  or an unusually large value of  $s_{\bar{X}}$ , or both.

### 11.3 ONE-SIDED TEST

Returning to our example relating to the flex life of tire cord, assume that on the basis of a random sample of 50 items we found

$$\bar{X} = 138.64 \text{ min}$$

$$s = 15.43 \text{ min}$$

Suppose that this time we wish to determine whether or not the cord has mean flex life less than 142 min. The test and the decision are given in outline form below. This outline may be considered standard for testing hypotheses.

**Hypotheses:**

$$H_0: \mu = 142 \text{ min}$$

$$H_1: \mu < 142 \text{ min}$$

**Criterion of Significance:**

$$\alpha = 0.05$$

**Rejection Region:**

With  $\nu = 49$ ,  $P(t | \nu) = P(t | 49) = 0.05$  and  $t_{0.05} = -1.68$ . Therefore, the region of rejection is  $t < -1.68$ .

**Estimated Standard Error:**

$$s_{\bar{X}} = \frac{s}{\sqrt{n}} = \frac{15.43}{\sqrt{50}} = 2.18 \text{ min}$$

**Test Statistic:**

$$t = \frac{\bar{X} - \mu_0}{s_{\bar{X}}} = \frac{138.64 - 142}{2.18} = -1.54$$

**Conclusion:**

$H_0$  is accepted since  $t > t_{0.05}$ ; i.e.,  $-1.54 > -1.68$ . Alternatively we may say that  $P(t | \nu) > \alpha$ ; i.e.,  $P(-1.54 | 49) \doteq 0.07$  which is greater than  $\alpha = 0.05$ .

## 11.4 CONFIDENCE LIMITS

When the population variance is not specified,  $t$  rather than  $z$  must be used in determining confidence limits. Otherwise the procedure is the same as was explained in Sec. 10.4. The appropriate formula for a  $100(1 - \alpha)$  percentage confidence interval is

$$\bar{X} - \mu_1 = \mu_2 - \bar{X} = t_{\alpha/2} s_{\bar{X}} \quad (11-4)$$

Thus, if  $n = 50$ ,  $\nu = 49$ ,  $\bar{X} = 138.64$  min,  $s = 15.43$  min, and  $1 - \alpha = 0.95$ ; then  $s_{\bar{X}} = 2.18$  min and  $t_{0.025} = 2.01$ . The confidence limits are

$$\mu_1 = 138.64 - 2.01(2.18) = 134.26 \text{ min}$$

$$\mu_2 = 138.64 + 2.01(2.18) = 143.02 \text{ min}$$

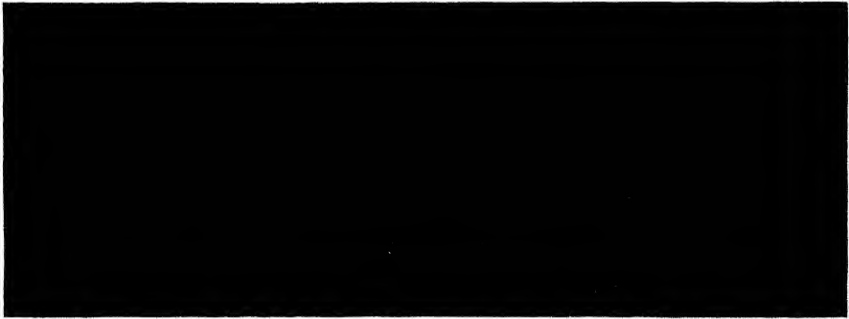
and the confidence interval is

$$\mu_2 - \mu_1 = 143.02 - 134.26 = 8.76 \text{ min}$$

The confidence interval is a little wider than we found when  $\sigma^2$  was known. This wider interval is a direct result of not having as much information to work with and is reflected in a value for  $t$  that is larger than the corresponding value of  $z$ . Another reason for the wider interval is that  $\sigma_{\bar{X}} = 2.12$  min, but  $s_{\bar{X}}$  was computed as 2.18 min. This, however, was an accident of sampling; sometimes  $s_{\bar{X}}$  will be larger than  $\sigma_{\bar{X}}$ , sometimes smaller.

Chart 11.3, which is analogous to Chart 10.10, differs from the latter in one important respect. Although  $\bar{X} - \mu_1$  always equals  $\mu_2 - \bar{X}$ ,  $\mu_1 - \mu_2$  varies from sample to sample. This variation occurs because  $s$ , and therefore  $s_{\bar{X}}$ , varies from sample to sample. As a consequence, the length of the vertical lines varies from sample to sample. There can be no uniform dotted lines

**CHART 11.3: 95 PER CENT CONFIDENCE LIMITS FOR MEAN, AND 40 SAMPLES, STANDARD DEVIATION UNKNOWN.**



above and below the solid line as in Chart 10.10. It is not necessarily the samples with the means deviating farthest from  $\mu$  that yield interval estimates failing to enclose  $\mu$ ; in some cases it is the sample with a small value of  $s$ .

## 11.5 HYPOTHESES CONCERNING DIFFERENCES BETWEEN TWO POPULATIONS

A comparison between two samples is frequently of greater interest than a comparison between a sample mean and a hypothetical value of  $\mu$ . In this section we will test the significance of the *difference between the means* of two independent samples as well as the significance of *mean difference* for matched samples.

**Independent Samples.** The data below summarize the pertinent results of drawing two random samples of size  $n = 50$ , one sample being drawn from a large lot of regular tire cord and the other from a large lot of Supertwist cord. The measurement is in terms of flex life in minutes. We wish

<i>Supertwist</i> (1)	<i>Regular</i> (2)
$\bar{X}_1 = 138.64$	$\bar{X}_2 = 87.66$
$\Sigma x_1^2 = 11,668$	$\Sigma x_2^2 = 9967$
$n_1 = 50$	$n_2 = 50$

to know if there is a significant difference between these two sample means. The null hypothesis is that the two sample means are computed from random samples drawn from the same population or from two populations having the same mean. Symbolically

$$H_0: \mu_1 - \mu_2 = 0$$

which we will test at  $\alpha = 0.10$ . We *assume* that each population is normally distributed with the same variance  $\sigma^2$ ; i.e., that  $\sigma_1^2 = \sigma_2^2 = \sigma^2$ . If we knew  $\sigma^2$ , we could test  $H_0$  by comparing  $\bar{X}_1 - \bar{X}_2$  with the standard error of the difference between two sample means, forming the  $z$  ratio.

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{X}_1 - \bar{X}_2}} \quad (11-5)$$

The quantity in the denominator of Eq. (11-5) is called the standard error of the difference between two sample means and, under the assumptions set out earlier, is given by the square root of

$$\sigma_{\bar{X}_1 - \bar{X}_2}^2 = \sigma_{\bar{X}_1}^2 + \sigma_{\bar{X}_2}^2 \quad (11-6)$$

The reasoning underlying Eq. (11-6) is as follows. If two populations are *independent*, the variance of the *sum or difference* of the elements in the populations is given by

$$\sigma_{\bar{X}_1 + \bar{X}_2}^2 = \sigma_{\bar{X}_1 - \bar{X}_2}^2 = \sigma_{\bar{X}_1}^2 + \sigma_{\bar{X}_2}^2 \quad (11-7)$$

The student can verify Eq. (11-7) in Problem 5. It should be remembered that Eq. (11-7) holds only in the case of independent populations.<sup>(3)</sup> Using similar reasoning, we add the individual variances of the sample means to get the variance of the difference between two sample means given by Eq. (11-6).

The problem is that rarely is  $\sigma^2$  known or specified, and we must substitute an estimate of  $\sigma^2$  that is a weighted average of the estimates of the population variance obtained from the two samples. This estimate of  $\sigma^2$  is given by<sup>(4)</sup>

$$s_W^2 = \frac{\nu_1 s_1^2 + \nu_2 s_2^2}{\nu_1 + \nu_2} \quad (11-8)$$

which simplifies to

$$s_W^2 = \frac{\sum x_1^2 + \sum x_2^2}{n_1 + n_2 - 2}$$

since  $s^2 = \sum x^2 / (n - 1)$ ,  $\nu_1 = n_1 - 1$ , and  $\nu_2 = n_2 - 1$ . The estimated variance of the difference between two sample means is given by

$$\sigma_{\bar{X}_1 - \bar{X}_2}^2 = \frac{s_W^2}{n_1} + \frac{s_W^2}{n_2} \quad (11-9)$$

---

<sup>(3)</sup> Equation (11-7) assumes that there is no correlation between the populations. If there is correlation

$$\sigma_{\bar{X}_1 - \bar{X}_2}^2 = \sigma_{\bar{X}_1}^2 + \sigma_{\bar{X}_2}^2 - 2\rho\sigma_{\bar{X}_1}\sigma_{\bar{X}_2}$$

where  $\rho$  is the population correlation coefficient.

<sup>(4)</sup> The expression  $s_W^2$  is equivalent to the "error" variance used in the analysis of variance.

Using the values calculated from the two tire cord samples, we have

$$s_W^2 = \frac{11,668 + 9967}{50 + 50 - 2} = 220.77$$

Then 
$$s_{\bar{X}_1 - \bar{X}_2}^2 = \frac{220.77}{50} + \frac{220.77}{50} = 8.831$$

and 
$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{8.831} = 2.97$$

The observed  $t$  statistic, analogous to the  $z$  statistic of Eq. (11-5), is

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}}$$

and for our example

$$t = \frac{138.64 - 87.66}{2.97} = 17.2$$

Each set of data contribute 49 degrees of freedom to the test. Under the null hypothesis the distribution of the differences between the means,  $\bar{X}_1 - \bar{X}_2$ , will have a mean of zero. Then for the two-sided alternative

$$H_1: \mu_1 - \mu_2 \neq 0$$

the rejection values for  $t$ , with  $\nu = 98$  and  $\alpha = 0.10$ , will be  $t_U = t_{0.05} \doteq 1.7$  and  $t_L = -t_U \doteq -1.7$ . Since  $t > t_U$ , i.e.,  $17.2 > 1.7$ , we reject  $H_0$  and conclude that there is a significant difference between the two means. The one-sided alternative would be tested by using  $t_\alpha$ .

In this particular problem the total number of degrees of freedom is large (almost 100), so the standardized normal distribution could have been substituted for the  $t$  distribution. The reader may verify that the decision would not have been altered by this substitution.

Let us now determine confidence limits for  $\mu_1 - \mu_2$ . To simplify the notation we let

$$\mu_1 - \mu_2 = \Delta\mu \quad \text{and} \quad \bar{X}_1 - \bar{X}_2 = D\bar{X}$$

Then,  $100(1 - \alpha)$  percent confidence limits may be established for  $\Delta\mu$  by solving

$$D\bar{X} - \Delta\mu_1 = \Delta\mu_2 - D\bar{X} = t_{\alpha/2} s_{\bar{X}_1 - \bar{X}_2} \quad (11-10)$$

where  $\nu = n_1 + n_2 - 2$ . For the present illustration the 90 percent confidence limits are

$$\Delta\mu_1 = 50.98 - 1.7(2.97) = 45.93 \text{ min}$$

$$\Delta\mu_2 = 50.98 + 1.7(2.97) = 56.03 \text{ min}$$

since  $D\bar{X} = 50.98$  min and  $t_{0.05} \doteq 1.7$  with 98 degrees of freedom.

**Matched Samples.** If some way can be found of matching each value of  $X_1$  with a related value of  $X_2$ , we can test the significance of *mean difference* rather than the significance of difference between means. This matching eliminates one source of variability in the data, leaving that which results from the lack of perfect association between the variables.

**TABLE 11.1: COMPRESSIVE STRENGTH OF 10 PAIRS OF TEST CUBES OF CONCRETE, AND COMPUTATIONS FOR TESTING SIGNIFICANCE OF MEAN DIFFERENCE**

<i>Batch</i>	<i>Treated</i> $X_1$	<i>Not treated</i> $X_2$	$D$ $X_1 - X_2$	$D^2$
1	309	293	16	256
2	318	311	7	49
3	317	284	33	1089
4	302	310	-8	64
5	315	305	10	100
6	296	291	5	25
7	319	301	18	324
8	285	279	6	36
9	303	295	8	64
10	290	289	1	1
Total	3054	2958	96	2008

Source: Ralph Allen Bradley, "Some Notes on the Theory and Applications of Rank Order Statistics, Part I," *Industrial Quality Control*, Vol. XI (February, 1955), p. 15.

For an illustration we shall use the data of Table 11.1, which are the compressive strength of 10 pairs of test cubes of concrete.

We use the symbol  $D$  to denote values of  $X_1 - X_2$  for the sample and  $\Delta$  to denote the same thing for the population. Since we are interested in discovering only whether the strength of the concrete is significantly increased by the treatment, it seems reasonable to use the one-sided test where  $\bar{\Delta}$  is the mean value of  $\Delta$  in the population. The distribution of  $D - \bar{\Delta}$  is assumed normal with mean of zero.

$$H_0: \bar{\Delta} = 0$$

$$H_1: \bar{\Delta} > 0$$

For the criterion of significance let us use  $\alpha = 0.025$  and proceed as if we were dealing with a single variable.<sup>(5)</sup>

$$\bar{D} = \frac{\sum D}{n} = \frac{96}{10} = 9.6 \text{ kilograms}$$

$$s_D = \sqrt{\frac{\sum D^2 - (\sum D)^2/n}{n-1}} = \sqrt{\frac{2008 - (96)^2/10}{9}} = 10.99 \text{ kilograms}$$

$$s_D = \frac{s_D}{\sqrt{n}} = \frac{10.99}{\sqrt{10}} = 3.476 \text{ kilograms}$$

$$t = \frac{\bar{D}}{s_D} = \frac{9.6}{3.476} = 2.76$$

<sup>(5)</sup> It can be shown that

$$s_D^2 = s_{X_1 - X_2}^2 = s_1^2 + s_2^2 - 2rs_1s_2$$

where  $r$  is the sample correlation coefficient. For the data of Table 11.1,  $s_1^2 = 149.156$ ;  $s_2^2 = 115.956$ ;  $r = 0.549$ ;  $s_D^2 = 120.71$ ;  $s_D = 10.99$ . An equivalent solution can be obtained by the very general method of analysis of variance. See Chapter 17.

With  $\nu = 9$  and  $\alpha = 0.025$ ,  $t_\alpha = t_{0.025} = 2.26$ . Since the computed value of  $t$  is greater than  $t_{0.025}$ , we reject  $H_0$  and conclude that the treatment significantly increases the strength of the concrete.

Confidence limits for  $\bar{D}$  may be established in the usual manner. For 100(1 -  $\alpha$ ) percent confidence limits we solve for  $\bar{D}_1$  and  $\bar{D}_2$ .

$$\bar{D} - \bar{D}_1 = \bar{D}_2 - \bar{D} = t_{\alpha/2} s_{\bar{D}} \quad (11-11)$$

where  $\nu = n - 1$ . For our example, the 95 percent limits are

$$\bar{D}_1 = 9.6 - 2.26(3.476) = 1.74 \text{ kilograms}$$

$$\bar{D}_2 = 9.6 + 2.26(3.476) = 17.46 \text{ kilograms}$$

since  $\bar{D} = 9.6$  kilograms and  $t_{0.025} = 2.26$ .

Although the pairing of  $X$  values reduces the standard error, it also reduces the number of degrees of freedom for the test. If we cannot pair the  $X$  values satisfactorily, a more powerful test may result from using independent samples. However, one should make the decision as to whether or not to pair before the sample is taken.

## PROBLEMS

1. Given the sample values in inches,

2, 2, 4, 4

test at  $\alpha = 0.10$

$$a. H_0: \mu = 4 \text{ inches} \quad b. H_0: \mu = 4 \text{ inches}$$

$$H_1: \mu \neq 4 \text{ inches} \quad H_1: \mu < 4 \text{ inches}$$

and set 99 percent confidence limits for  $\mu$ .

2. Eight secretaries were selected at random from those working in a large insurance company. Four of the women were given a typing assignment to perform on one kind of electric typewriter, and the other four were given the same assignment to perform on another kind of electric typewriter. None of the women had ever had experience with either make of typewriter or with the assignment. The error adjusted time needed to complete the assignment (in minutes) is given below.

Group 1 (min)	Group 2 (min)
3	1
2	2
3	1
4	2

Assume that  $\sigma_1^2 = \sigma_2^2 = \sigma^2$  and that both populations are independent and normally distributed. Test, at  $\alpha = 0.05$ ,

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 \neq 0$$

If you were in charge of purchasing typewriters, would you prefer the machine used by Group 2 over that used by Group 1 if the machines were equally expensive? How might differences in the cost of the two types of machines affect your choice of  $\alpha$ ? Set 95 percent confidence limits for  $\mu_1 - \mu_2$ .

3. Four men were placed on a carefully supervised diet. Their weights before and after the diet are given below.

Man	Weight before diet ( $X_1$ )	Weight after diet ( $X_2$ )
1	225 lbs	220 lbs
2	250	190
3	210	200
4	275	250

Test

$$H_0: \bar{\Delta} = 0 \quad \text{at} \quad \alpha = 0.05$$

$$H_1: \bar{\Delta} < 0$$

Set 95 percent confidence limits for  $\bar{\Delta}$ .

4. Let  $z^* = (\bar{X} - \mu)/\sigma$  and  $z = (X - \bar{X})/\sigma$ . Assuming an infinite population, show that Eq. (11-3) can be written as

$$t = \frac{z^* \sqrt{n}}{\sqrt{\sum z^2 / (n-1)}}$$

Also show that if  $R$  is the range of a random sample of size  $n = 2$ , Eq. (11-3) may be written as

$$t = \frac{2(\bar{X} - \mu)}{R}$$

5. The following pairs of  $X$  values are uncorrelated (as will be shown in Chapter 15). For the present, assuming that the series represent finite populations, show that  $\sigma_{X_1+X_2}^2 = \sigma_{X_1}^2 + \sigma_{X_2}^2 = \sigma_{X_1-X_2}^2$ .

$X_1$	$X_2$
2	1
1	2
3	2
2	3



# 12

## Tests of Hypotheses and Confidence Limits for Proportions and Standard Deviations

This chapter extends the discussion of hypothesis testing and the setting of confidence limits to proportions and standard deviations. Before undertaking this task, we should perhaps review and extend some results previously obtained with respect to proportions.

### 12.1 A REVIEW AND EXTENSION OF SOME PREVIOUS RESULTS CONCERNING PROPORTIONS

The following statistics and parameters, all of which have been previously defined, will be used in the next few sections.

<i>Measure</i>	<i>Statistic</i>	<i>Parameter</i>
Number of "defective" items	$d$	$D$
Number of "effective" items	$g = n - d$	$G = N - D$
Proportion "defective"	$p = d/n$	$P = D/N$
Proportion "effective"	$q = g/n$	$Q = G/N$

We saw in Sec. 8.2 that

$$E(p) = P \quad (12-1)$$

and

$$E(d) = nP \quad (12-2)$$

Furthermore, we recall that Eq. (12-2) is the expected value of the binomial probability distribution, and it was shown in Sec. 8.5 that for an infinite population the variance of the statistic  $p$  is given by

$$\sigma_p^2 = \frac{PQ}{n} \quad (12-3)$$

Also, the variance of the statistic  $d$ , which is the variance of the binomial probability distribution, is

$$\sigma_d^2 = nPQ \quad (12-4)$$

For finite populations the variances of  $p$  and  $d$  will be subject to a finite population correction factor just as is the variance of the mean,  $\sigma_{\bar{x}}^2$ . For *finite populations*<sup>(1)</sup>

$$\sigma_p^2 = \frac{PQ}{n} \left( \frac{N-n}{N-1} \right) \quad (12-5)$$

and 
$$\sigma_d^2 = nPQ \left( \frac{N-n}{N-1} \right) \quad (12-6)$$

The term  $(N-n)/(N-1)$  may be called the finite population correction factor for proportions. Further, just as  $E(d)$  and  $\sigma_d^2$ , as given by Eqs. (12-2) and (12-4), are the mean and variance of the binomial distribution when  $N$  is infinite, we shall see in a later section that *when  $N$  is finite*,  $E(d)$  and  $\sigma_d^2$ , as given by Eqs. (12-2) and (12-6), are the mean and variance of the *hypergeometric distribution*.

We mention two final points of review. First, the normal distribution is the limit of the binomial density as  $n$  approaches infinity, with  $P$  constant. This fact, discussed in Sec. 7.1, is extremely important, since it allows us to approximate the binomial distribution using the normal distribution for large values of  $n$  at a considerable saving of time and effort. We know from Sec. 7.1 that the "speed" at which the binomial distribution approaches the normal distribution depends upon the skewness of the binomial distribution, which, in turn, depends upon the relationship between  $P$  and  $Q$ . It turns out that the normal approximation to the binomial is accurate enough for most practical work when  $nP \geq 25$ . Since  $nP$  is the mean of the binomial distribution, we are saying that the approximation is accurate whenever  $E(d) \geq 25$ , regardless of the value of  $P$ .

Second, it should be recalled that the Poisson distribution is the limit of the binomial distribution as  $n$  approaches infinity, with  $nP$  constant and that the Poisson may be substituted for the binomial with negligible error and considerable saving of effort when  $n \geq 10$  and  $P \leq 0.1$ .

<sup>(1)</sup> A discussion of Eqs. (12-5) and (12-6) is in an appendix to this chapter. In some texts  $(N-n)/(N-1)$  is also used as the finite population correction factor for the variance of the mean. Such usage implies a different definition of the variance of a finite population than the one adopted in this text.

## 12.2 HYPOTHESES FOR PROPORTIONS, INFINITE POPULATIONS

**Use of the Binomial Distribution.** A lot (assumed to be very large) is considered satisfactory if it is not more than 2 percent defective. We select a random sample of 100 items and find 6 defective. We are willing to subject the producer to a risk of not greater than 10 percent of rejecting a lot when the lot quality  $P$  is good. Should the lot be accepted or rejected?

*Hypotheses:*

$H_0: P = 0.02$  (lot quality "good")

$H_1: P > 0.02$  (lot quality "bad")

*Criterion of Significance (Producer's Risk):*

$\alpha = 0.10$

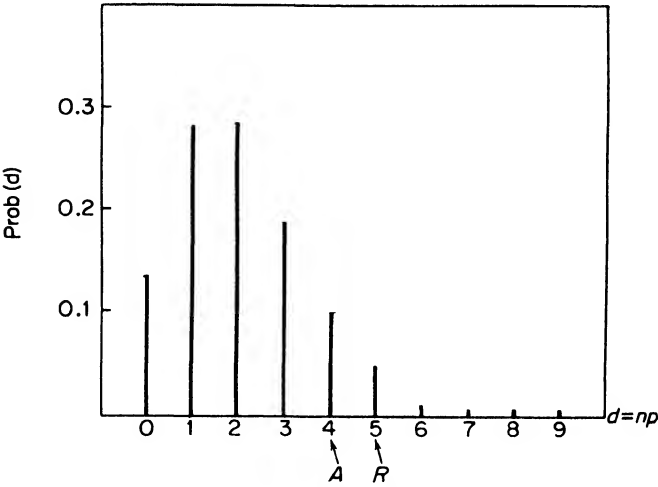
Since  $nP = 100(0.02) = 2$ , it would not be sufficiently accurate to use normal probabilities by the rule of thumb given in the last section. Using Eq. (6-17), or a binomial table, we construct the binomial probability distribution given in Table 12.1. The distribution is plotted in Chart 12.1.

TABLE 12.1: BINOMIAL PROBABILITY DISTRIBUTION,  $P = 0.02$ ;  $n = 100$

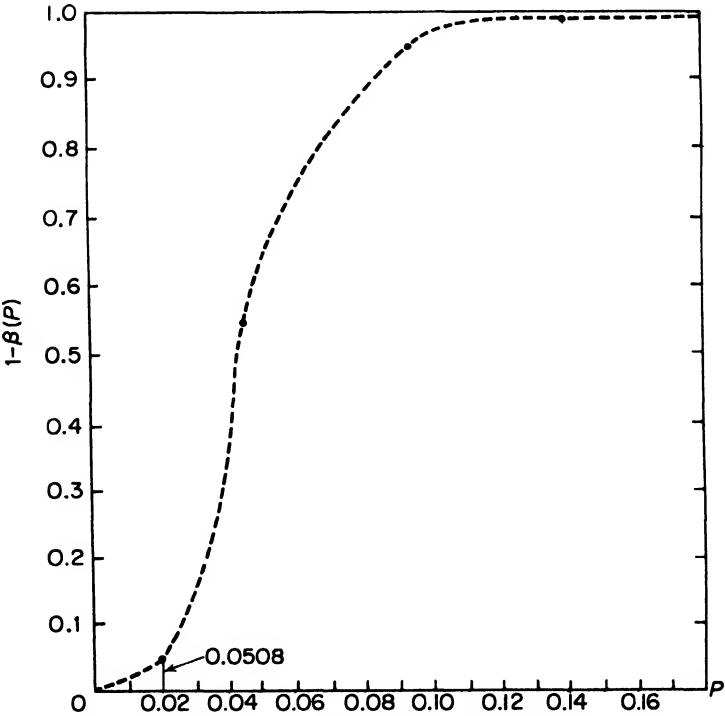
Sample number of defectives $d$	Sample proportion defective $p$	Prob ( $d$ ), or Prob ( $p$ )	Cumulative Prob ( $d$ ) $Q(d)$
0	0.00	0.1326	1.0000
1	0.01	0.2707	0.8674
2	0.02	0.2734	0.5967
3	0.03	0.1823	0.3233
4 = $A$	0.04	0.0902	0.1410 $> \alpha = 0.10$
5 = $R$	0.05	0.0353	0.0508 $< \alpha = 0.10$
6	0.06	0.0114	0.0155
7	0.07	0.0032	0.0041
8	0.08	0.0007	0.0009
9	0.09	0.0002	0.0002

We wish to establish a number of defective items  $R$ , called a *rejection number*, such that  $R$  bounds from the left an area in the right tail of the binomial probability distribution *not greater* than  $\alpha$ . Notice that we are almost never able to bound exactly  $\alpha$  percent in the right tail of the binomial distribution, since the distribution is discrete. Also, since  $\alpha$  is the *maximum* probability of a type I error that we wish to allow, the area in the right tail that is bounded on the left by  $R$  cannot exceed  $\alpha$ . Cumulating the binomial distribution from the right tail, we see that  $d = 5$  is the largest value of  $d$

**CHART 12.1: BINOMIAL DISTRIBUTION AND POWER CURVE ASSOCIATED WITH ONE-SIDED TEST OF A HYPOTHESIS FOR PROPORTIONS  $n = 100$ ;  $P = 0.02$ .**



(a) Binomial distribution



(b) Power curve

that satisfies our criterion, and the rejection number is  $R = 5$ ; the acceptance number is  $A = 4$ . Thus, if 4 or fewer defective items are found, the null hypothesis should be accepted; if 5 or more defective items are found, the null hypothesis should be rejected.

**Conclusion:**

Since we found 6 defective items, we reject the null hypothesis and reject the lot.

The power curve for this test is fairly easy to construct. It is the probability of rejecting the lot as a function of  $P$  with the rejection number fixed. We already have three points for the power curve:

When  $P = 0.00$ ,  $1 - \beta(P) = 0$  (the sample will contain effective items only).

When  $P = 0.02$ ,  $1 - \beta(P) = 0.0508$  (from Table 12.1).

When  $P = 1.00$ ,  $1 - \beta(P) = 1.000$  (the sample will contain defective items only).

Three more intermediate points will allow us to plot the power curve with reasonable accuracy:

When  $P = 0.05$ ,  $1 - \text{Prob}(d < 4) = 0.5640$ .

When  $P = 0.10$ ,  $1 - \text{Prob}(d < 4) = 0.9763$ .

When  $P = 0.15$ ,  $1 - \text{Prob}(d < 4) = 0.9996$ .

These points are obtained in the same manner as the point for  $P = 0.02$  and are plotted in the lower section of Chart 12.1.<sup>(2)</sup>

**Use of the Poisson Distribution.** Continuing with the same example, we see that since  $n = 100 > 10$  and  $P = 0.02 < 0.1$ , the Poisson distribution should give us a good approximation to the binomial. Table 12.2 shows the values of the relevant Poisson distribution. It is clear that we find the same acceptance and rejection number as we did using the more laborious binomial distribution.

**Use of the Normal Distribution.** There are various "cola" drinks on the market, and the question is often raised as to whether it is possible for a person to identify a given brand by taste. A test was run by one of the authors concerning R-C Cola and Coca-Cola (registered trade marks). Each of 44 persons was given a drink of R-C Cola and a drink of Coca-Cola

<sup>(2)</sup> When control over both type I and type II errors is desired, one may specify two points on the power curve just as was done in the appendix to Chapter 10. For a one-sided test, the first point on the power curve would specify a given  $P_1$  and  $\alpha$  (producer's risk). A good lot would be one of quality  $P_1$ . The second point would specify  $\bar{P}_2$  and  $\beta_2$  (the probability of accepting the lot when the lot is of quality  $\bar{P}_2$ , or bad). The error probability  $\beta_2$  is often called *consumer's risk* in quality control contexts.

TABLE 12.2: POISSON DISTRIBUTION FOR  $nP = 2$ 

$d$	$Prob(d)$	$Cumulative$ $Prob(d), Q(d)$
0	0.1353	1.0000
1	0.2707	0.8646
2	0.2707	0.5939
3	0.1804	0.3232
4 = $A$	0.0902	0.1428 $> \alpha = 0.10$
5 = $R$	0.0361	0.0526 $< \alpha = 0.10$
6	0.0120	0.0165
7	0.0034	0.0045
8	0.0009	0.0011
9	0.0002	0.0002

and asked to state which was which. Thirty-four answered correctly, and 10 answered incorrectly. If the two drinks were identical, half of the subjects would be expected to distinguish between them by chance.

Now  $nP = 44(0.5) = 22$ , which is almost 25, and so the normal approximation to the binomial should be reasonably accurate. We proceed in the conventional manner.

*Hypotheses:*

$$H_0: P = 0.5$$

$$H_1: P > 0.5$$

*Criterion of Significance:*

$$\alpha = 0.05$$

*Rejection Region:*

From Appendix 3, we obtain  $z_{0.05} = 1.645$ . The null hypothesis will be rejected if  $z \geq 1.645$ .

*Standard Error of  $d$ :*

$$\sigma_d = \sqrt{nPQ} = \sqrt{44(0.5)(0.5)} = 3.317$$

*Test Statistic:<sup>(3)</sup>*

$$z = \frac{d - nP}{\sigma_d} = \frac{34 - 22}{3.317} = 3.62$$

<sup>(3)</sup> We might, if it were more convenient, have used  $z = (p - P)/\sigma_p = (0.773 - 0.50)/0.0754 = 3.62$ , where  $\sigma_p = \sqrt{PQ/n}$ . Also, with small samples sometimes one corrects for continuity.

$$z = \frac{|d - nP| - 0.5}{\sigma_d} = \frac{|p - P| - 0.5}{\sigma_p}$$

Without use of the correction,  $z = 3.62$  and  $Q(z) = 0.00015$ . The true probability, when binomial tables are used, is 0.00019. Using the correction factor, we have  $z = 3.47$  and  $Q(z) = 0.00026$ . In the present case the correction factor "overcorrects," and the uncorrected  $z$  value is more accurate than the corrected one.

**Conclusion:**

$H_0$  is rejected, since  $z > z_\alpha$ . We conclude that more than 50 percent of the population is able to distinguish between the two drinks. In other words, with  $z = 3.62$ ,  $Q(z) = 0.00015$ , which is less than  $\alpha$ .

## 12.3 HYPOTHESES FOR PROPORTIONS, FINITE POPULATIONS

**Use of Hypergeometric Distribution.** As previously noted in Sec. 12.1, when  $N$  is finite, the statistic  $d$  is distributed according to the hypergeometric distribution, which is given by

$$\text{Prob}(d) = \frac{\binom{D}{d} \binom{G}{g}}{\binom{N}{n}} = \frac{\binom{N \cdot P}{n \cdot p} \binom{N \cdot Q}{n \cdot q}}{\binom{N}{n}} \quad (12-7)$$

The distribution has mean given by Eq. (12-2) and variance given by Eq. (12-6). The binomial distribution may be thought of as the limit of the hypergeometric distribution as  $N$  approaches infinity,  $n$  remaining constant. Therefore, the binomial distribution may be substituted for the hypergeometric distribution when the ratio  $n/N$  is small. Experience suggests that the binomial may be substituted for the hypergeometric when  $n/N < 0.10$ . We will illustrate the use and computation of the hypergeometric distribution with the following problem.

A random sample of 3 items is taken without replacement from a lot of 50 items. One of these items is found to be defective. We are willing to subject the producer to a risk of not greater than 5 percent of rejecting the lot when the lot is good. A good lot is one with not more than 10 percent defective items. Should the lot be accepted or rejected?

**Hypotheses:**

$$H_0: P = 0.10$$

$$H_1: P > 0.10$$

**Criterion of Significance (producer's risk):**

$$\alpha = 0.05$$

**Rejection Number:**

Table 12.3 tabulates the hypergeometric distribution associated with this problem. Following the same procedure as outlined in the last section, we find the acceptance number to be  $A = 1$  and the rejection number,  $R = 2$ .

TABLE 12.3: HYPERGEOMETRIC DISTRIBUTION

 $N = 50; G = 45; n = 3$ 

(1)	(2)	(3)	(4)	(5)	(6*)	(7)
$d$	$g$	$\binom{5}{d}$	$\binom{45}{g}$	Col. (3) · Col. (4)	$\text{Prob } (d) =$ $\text{Col. (5)} / \binom{N}{n}$	<i>Cumulative</i> <i>Prob (d), Q(d)</i>
0	3	1	14,190	14,190	0.7240	1.0000
1 = A	2	5	990	4,950	0.2526	$0.2761 > \alpha = 0.05$
2 = R	1	10	45	450	0.0230	$0.0235 < \alpha = 0.05$
3	0	10	1	10	0.0005	0.0005

$$* \binom{N}{n} = \binom{50}{3} = 19,600.$$

**Conclusion:**

Since we found only one defective item, we accept the lot.

**Use of the Binomial Distribution.** Since  $n/N = \frac{3}{50} = 0.06$ , which is less than 0.10, we may substitute the binomial distribution for the hypergeometric distribution with considerable saving of effort and negligible loss of accuracy. Table 12.4 shows the binomial distribution associated with

TABLE 12.4: BINOMIAL DISTRIBUTION

 $n = 3; P = 0.10$ 

$d$	<i>Prob (d)</i>	<i>Cumulative</i> <i>Prob (d), Q(d)</i>
0	0.7290	1.000
1 = A	0.2430	$0.2710 > \alpha = 0.05$
2 = R	0.0270	$0.0280 < \alpha = 0.05$
3	0.0010	0.0010

this problem, and we see that we receive the same acceptance and rejection numbers as before.

## 12.4 CONFIDENCE LIMITS OF A PROPORTION

When dealing with sample means, we found it useful to ascertain the confidence limits of  $\mu$  from a knowledge of the information given by one sample. Similarly, it is sometimes important to determine the confidence limits of  $P$  when only  $p$  and  $n$  are known.

The Chicago, Milwaukee, and Saint Paul and Pacific Railway tested various sorts of woods and preservatives for railroad ties. One lot of 50 red



oak ties treated with 20 percent creosote and 80 percent zinc chloride was examined after 20 years of use, and 19, or 38 percent of them, were found still to be good. Our sample values are  $n = 50$  and  $p = 0.38$ . Let us ascertain the 95 percent confidence limits of  $P$ . We want first to ascertain  $P_1$ , less than  $p$ , so located that  $p$  cuts off the upper 2.5 percent tail of the distribution of sample  $p$ 's around  $P_1$ , and, second, to determine  $P_2$ , greater than  $p$ , so located that  $p$  cuts off the lower 2.5 percent tail of the distribution of sample  $p$ 's around  $P_2$ .

### Use of Binomial Distribution (Clopper-Pearson Charts).

Charts for estimating the confidence limits of  $P$ , based on the binomial distribution, are shown in Appendix 8 for 95 and 99 percent intervals. To illustrate the use of these charts for our problem, erect a perpendicular line at  $p = 0.38$  on the chart relevant for 95 percent limits. Where this line intersects the two curves for  $n = 50$ , draw two horizontal lines. Notice that the points of intersection must be estimated. There are curves for  $n = 40$  and for  $n = 60$ , but not  $n = 50$ . The two horizontal lines intersect the  $P$  axis of the chart at  $P_1 \doteq 0.25$  and  $P_2 \doteq 0.53$ . These values of  $P_1$  and  $P_2$  are the approximate 95 percent confidence limits. Notice that the limits are not symmetrical about  $p = 0.38$  because of the lack of symmetry in the binomial distribution itself when  $P \neq 0.5$  and because the binomial changes shape and dispersion as  $P$  changes. Also, the closer  $p$  is to 0.5, the greater the distance between  $P_2$  and  $P_1$ . This is easy to understand when one realizes that  $\sigma_p$  is at a maximum when  $P = 0.5$ .

**Use of Binomial Distribution (Direct Mathematical Solution).** A mathematical solution, equivalent to the graphic one above but usually more accurate, is as follows: Find the largest value of  $P_1$ , smaller than  $p$ , and such that

$$\text{Prob}(d > 19) < 0.025$$

$$\text{i.e.,} \quad \sum_{d=19}^{50} \binom{n}{d} P_1^d Q_1^{n-d} < 0.025$$

Using Romig's *Tables*,<sup>(4)</sup> we see that it appears that when  $P_1 \doteq 0.246$  this equation is satisfied. Next, find the smallest value of  $P_2$ , larger than  $p$ , and such that

$$\text{Prob}(d < 19) < 0.025$$

$$\text{i.e.,} \quad \sum_{d=0}^{19} \binom{n}{d} P_2^d Q_2^{n-d} < 0.025$$

<sup>(4)</sup> Harry G. Romig, *50-100 Binomial Tables* (New York: John Wiley & Sons, Inc. 1953).

When we try to determine  $P_2$  by use of Romig's *Tables*, we run into a slight snag. It soon becomes evident that  $P_2 > 0.5$ , but Romig's *Tables* only go to  $P = 0.5$ . This is not a very serious snag, however. All we need to do is to interchange  $P$  and  $Q$ , and also  $d$  and  $n - d$ . In our equation for finding  $P_2$  we therefore substitute 31 for 19, and interchange  $P_2$  and  $Q_2$  to find

$$\text{Prob } (d \leq 31) \leq 0.025$$

i.e., 
$$\sum_{d=31}^{50} \binom{n}{d} Q_2^d P_2^{n-d} \leq 0.025$$

Romig's *Tables* show that the largest value of  $Q_2 = 0.471$ . Therefore,  $P_2 = 1.0 - 0.471 = 0.529$ .

The accuracy of the mathematical solution is limited only by the extent of one's willingness to experiment with trial values of  $P_1$  and  $P_2$ .

**Use of Normal Distribution.** Although we do not yet know the value of  $nP_1$  or  $nP_2$ , the value of  $np = d = 19$  is sufficiently close to 25 to justify the use of the normal distribution as an approximate solution. The equation for confidence limits for  $P$  parallels those used in previous sections.

$$p - P_1 = P_2 - p = z_{\alpha/2} \sigma_p \quad (12-8)$$

But since  $\sigma_p = \sqrt{(P - P^2)/n}$ , we may drop the subscripts on  $P$ , and write Eq. (12-8) as

$$(np^2) - (z_{\alpha/2}^2 + 2np)P + (z_{\alpha/2}^2 + n)P^2 = 0 \quad (12-9)$$

Substituting the values of  $n$ ,  $p$ , and  $z_{\alpha/2}$  given in our problem into Eq. (12-9) and simplifying, we get

$$7.220 - 41.8416P + 53.8416P^2 = 0$$

This is a quadratic equation of the type

$$a + bX + cX^2 = 0$$

and can be solved for real and unequal roots by use of

$$X = \frac{-b \pm \sqrt{b^2 - 4ac}}{2c}$$

if  $b^2 - 4ac$  is positive. For our problem

$$P = \frac{41.8416 \pm \sqrt{(41.8416)^2 - 4(7.220)(53.8416)}}{2(53.8416)}$$

and the solutions are

$$P_1 = 0.259 \quad \text{and} \quad P_2 = 0.519$$

## 12.5 HYPOTHESES CONCERNING DIFFERENCES BETWEEN PROPORTIONS

The results of a survey conducted in two rural counties for a tire company called "Superior" are given below.

	County 1	County 2
Number of persons sampled	$n_1 = 954$	$n_2 = 770$
Persons who plan to buy Superior tires	$d_1 = 259$	$d_2 = 206$
Percentage who plan to buy Superior tires	$\frac{d_1}{n_1} = p_1 = 27.15$	$\frac{d_2}{n_2} = p_2 = 26.75$

We consider that the dealer in county one is the more aggressive, and we wish to test the hypothesis that his aggressiveness makes a difference. That is, it is desired to know if the difference between the two proportions is significant.

The test follows the general outline given in Sec. 11.5 for testing the difference between two means when the samples are independent.

*Hypotheses:*

$$H_0: P_1 - P_2 = 0$$

$$H_1: P_1 - P_2 \neq 0$$

*Criterion of Significance:*

$$\alpha = 0.10$$

*Estimate of P:*

This is the weighted mean of  $p_1$  and  $p_2$ .

$$\bar{p} = \frac{d_1 + d_2}{n_1 + n_2} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{259 + 206}{954 + 770} = 0.2697$$

*Variance of Difference between  $p_1$  and  $p_2$ :*

$$s_{p_1 - p_2}^2 = \frac{\bar{p}\bar{q}}{n_1} + \frac{\bar{p}\bar{q}}{n_2} = \frac{0.19696}{954} + \frac{0.19696}{770} = 0.0004622$$

where  $\bar{q} = 1 - \bar{p}$ . Then

$$s_{p_1 - p_2} = \sqrt{0.0004622} = 0.02150$$

*Rejection Region:*

Since we are dealing with a large sample, we may consider that  $z$ , the ratio of the difference to its estimated standard error, is normally distributed.<sup>(5)</sup> Therefore, for a two-sided test the rejection values of  $z$  are  $z_L = -1.645$  and  $z_U = 1.645$ .

<sup>(5)</sup> An equivalent test, using a  $2 \times 2$  contingency table, is explained in Chapter 17.

**Test Statistic:**

$$z = \frac{p_1 - p_2}{s_{p_1 - p_2}} = \frac{0.2715 - 0.2675}{0.0215} = 0.19$$

**Conclusion:**

Since  $z$  is between  $z_L$  and  $z_U$ , we do not reject  $H_0$ .

## 12.6 TESTS OF HYPOTHESES AND CONFIDENCE LIMITS FOR THE STANDARD DEVIATION

In order to handle the problems in this section we introduce a new statistic,  $\chi^2$  (chi square), which may be defined as the sums of squares of  $\nu$  independent standardized normal deviates.

$$\chi^2 = \sum_1^{\nu} \left( \frac{X - \mu}{\sigma} \right)^2 = \sum_1^{\nu} z^2 \quad (12-10)$$

The statistic given in Eq. (12-10) is distributed according to the chi square distribution with  $\nu$  degrees of freedom. It follows that if the  $X$  values are normally distributed in the population

$$\chi^2 = \frac{\sum_1^n (X - \bar{X})^2}{\sigma^2} = \frac{\nu s^2}{\sigma^2} \quad (12-11)$$

has the chi square distribution with  $\nu = n - 1$  degrees of freedom and cannot be negative but may be indefinitely large.<sup>(6)</sup>

Values of  $\chi^2_Q$  for various levels of  $Q(\chi^2 | \nu)$  are tabulated in Appendix 6. Since a  $\chi^2$  distribution, like a  $t$  distribution, is a function of  $\nu$ , the tabulation pattern follows that used for the  $t$  distribution.

**Tests of Hypotheses:**

Suppose that we desire to test

$$H_0: \sigma = 3.5 \text{ lb}$$

$$H_1: \sigma \neq 3.5 \text{ lb}$$

at  $\alpha = 0.02$ . A random sample of size  $n = 15$  is taken from a process, and it is found that  $s = 6.0$  lb. Should we reject  $H_0$ , given  $\alpha$ ? Following usual procedures, we wish to establish an upper and lower rejection value for  $\chi^2$ . Thus, we desire two values of  $\chi^2$ , a value  $\chi^2_L$  such that  $P(\chi^2_L | \nu) = \alpha/2 = 0.01$  and a value  $\chi^2_U$  such that  $Q(\chi^2_U | \nu) = \alpha/2 = 0.01$ . Since a chi square distribution is not symmetrical, we must find each of the values individually in Appendix

<sup>(6)</sup> One degree of freedom is lost through the use of the sample mean as an estimator of the population mean.

6. Using this appendix with  $\nu = n - 1 = 14$ , we locate  $Q(\chi^2 | \nu) = 0.01$  and find  $\chi^2_U = 29.141$ . In a similar manner  $P(\chi^2 | \nu) = 1 - Q(\chi^2 | \nu)$  and we locate  $Q(\chi^2 | \nu) = 0.99$ . Therefore, we see that  $\chi^2_L = 4.660$ .

The observed value of chi square

$$\chi^2 = \frac{\nu s^2}{\sigma_0^2} = \frac{14(6)^2}{(3.5)^2} = 41.14$$

exceeds the upper rejection value  $\chi^2_U = 29.141$ , and we reject  $H_0$ . Rejection values for  $s^2$  can be determined by use of the relationships

$$s^2_U = \frac{\sigma^2(\chi^2_Q)}{\nu}$$

$$\text{and } s^2_L = \frac{\sigma^2(\chi^2_P)}{\nu}$$

Also, a one-sided test can be made, using  $Q(\chi^2 | \nu) = \alpha$  or  $P(\chi^2 | \nu) = \alpha$ , depending upon the direction of the test.<sup>(7)</sup>

**Confidence Limits.** Confidence limits may be set for  $\sigma$  by using Eq. (12-11). The  $100(1 - \alpha)$  percent limits are found by rearranging Eq. (12-11) and solving

$$\sigma_1 = \sqrt{\frac{\nu s^2}{\chi^2_{\alpha/2}}} \quad (12-12)$$

for the lower limits with  $\nu = n - 1$  and

$$\sigma_2 = \sqrt{\frac{\nu s^2}{\chi^2_{1-\alpha/2}}} \quad (12-13)$$

for the upper limits with  $\nu = n - 1$ . The confidence limits will not be symmetrical about  $s$ , since the given  $\chi^2$  distribution will not be symmetrical.<sup>(8)</sup>

## 12.7 HYPOTHESIS THAT TWO POPULATIONS HAVE THE SAME STANDARD DEVIATION

**The F Distribution.** The  $F$  distribution applies to the ratio of two independently distributed quantities, each following the chi square distribution with  $\nu_1$  and  $\nu_2$  degrees of freedom, respectively. The  $F$  (or variance)

<sup>(7)</sup> In a similar manner,  $(SD)^2$  could be used for the test, since

$$\chi^2 = \frac{n(SD)^2}{\sigma^2}$$

with  $n - 1$  degrees of freedom. Also, a method of evaluating the power function for this test can be deduced from Dudley J. Cowden, *Statistical Methods in Quality Control* (Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1957), pp. 113ff.

<sup>(8)</sup> Confidence limits using  $SD$  may be set by replacing  $\nu s^2$  with  $n(SD)^2$  in Eqs. (12-12) and (12-13).

ratio may be given by

$$F = \frac{\chi_1^2/\nu_1}{\chi_2^2/\nu_2} \quad (12-14)$$

and follows the  $F$  distribution with  $\nu_1$  and  $\nu_2$  degrees of freedom. Using Eq. (12-11), we may write Eq. (12-14) for two independent estimates of the population variance:

$$F = \frac{s_1^2}{s_2^2} \quad (12-15)$$

Since the sample variance cannot be less than zero,  $F$  cannot be negative but may be indefinitely large. Appendix 5 tabulates values of  $F_\alpha$  for various levels of  $Q(F|\nu_1, \nu_2)$ . Only the upper probability points are tabulated, since these are the ones most often useful.<sup>(9)</sup>

**Test of Hypothesis.** To illustrate the use of the  $F$  ratio, suppose that we have taken two independent random samples of size  $n = 50$ , one being drawn at random from a large lot of Supertwist cord and the other being drawn at random from a large lot of regular cord. The two sample variances are found to be

$$s_1^2 = 238.12 \text{ (for Supertwist)}$$

$$s_2^2 = 203.41 \text{ (for regular)}$$

Let us perform a one-sided test of equality of variances at the 0.05 level.

*Hypotheses:*

$$H_0: \sigma_1^2 - \sigma_2^2 = 0$$

$$H_1: \sigma_1^2 - \sigma_2^2 > 0$$

*Criterion of Significance:*

$$\alpha = 0.05$$

*Rejection Region:*

For  $s_1^2$ ,  $\nu_1 = n - 1 = 49$  and for  $s_2^2$ ,  $\nu_2 = n - 1 = 49$ . Entering Appendix 5, we find that at  $\alpha = 0.05$ ,  $\nu_1 = \nu_2 = 49$ , the upper rejection value for  $F$  is approximately 1.6. Thus, if the observed value of  $F$  exceeds 1.6 we will reject  $H_0$ .

<sup>(9)</sup> The lower probability is the reciprocal of  $F_\alpha$  with  $\nu_1$  and  $\nu_2$  interchanged. Thus, if  $\nu_1 = 5$  and  $\nu_2 = 8$ , the 0.05 upper probability point is  $F_\alpha = 3.69$ . If  $\nu_1 = 8$  and  $\nu_2 = 5$ , the 0.05 upper probability point  $F_\alpha = 4.82$ , and the 0.05 lower probability point  $F_\alpha$  is  $1/3.69 = 0.271$ .

**Test Statistic:**

$$F = \frac{238.12}{203.41} = 1.171$$

**Conclusion:**

The observed value of  $F$  does not fall into the rejection region. We do not reject the null hypothesis.<sup>(10)</sup> The probability of obtaining a value of  $s_1^2/s_2^2$  as large as, or larger than, the observed value is considerably more than 0.05.

When  $\nu_1 = \nu_2$ , and the probability is 0.05 for a given value of  $s_1^2/s_2^2$ , then the probability is 0.10 of obtaining a value of  $s_1^2/s_2^2$  as large as or larger than that obtained or a value of  $s_1^2/s_2^2$  as small as or smaller than the reciprocal of that obtained.

## PROBLEMS

1. If  $p = 0.5$  and  $n = 20$ , use Appendix 8 to set 95 and 99 percent confidence limits for  $P$ .
2. Of 500 students interviewed about a political issue, 300 were in favor of United States policy, and 200 were opposed. Test

$$H_0: P = 0.5$$

$$H_1: P > 0.5$$

at  $\alpha = 0.05$ . What can you say about the sentiment of the entire student body if the sample is assumed random?

3. In a box of 40 glasses 5 were sampled at random, and one of these 5 was broken. The box will be accepted if not more than 5 percent of all glasses are broken. Let  $\alpha = 0.10$  and decide whether the box should be accepted.

4. After the interviews described in Problem 2 were made, 700 students were interviewed on another campus. 350 were in favor of United States policy and 350 were opposed. Conduct a two-sided test of the hypothesis that there is no difference between the proportion of favorable responses in the two schools. Let  $\alpha = 0.10$ .

5. What assumption made in Sec. 11.5 can be tested by use of the discussion in Sec. 12.7?

<sup>(10)</sup> A discussion of the power function for the " $F$  test" along with tables for selected values of  $\alpha$  is given in: Moti Lal Tiku, "Tables of the Power of the  $F$ -test," *Journal of the American Statistical Association*, June, 1967, pp. 525-539.

### APPENDIX: Discussion of the Variance of $d$ and $p$ , Finite Population

For a finite population the variance of the population may be calculated by use of

$$\sigma_X^2 = \frac{\sum X^2 - (\sum X)^2/N}{N-1} = \left[ \frac{\sum X^2}{N} - \left( \frac{\sum X}{N} \right)^2 \right] \left( \frac{N}{N-1} \right) \quad (\text{A12-1})$$

and the variance of the mean is

$$\sigma_{\bar{X}}^2 = \frac{\sigma_X^2}{n} \left( 1 - \frac{n}{N} \right) \quad (\text{A12-2})$$

In the case of proportions the  $X$  values may take on only the values 0 and 1, and we note that in this case

$$\sum X^2 = \sum X \quad X = 0, 1 \quad (\text{A12-3})$$

Using Eq. (A12-3), we substitute  $\sum X$  for  $\sum X^2$  in the second form of Eq. (A12-1), obtaining

$$\sigma_X^2 = \left[ \frac{\sum X}{N} - \left( \frac{\sum X}{N} \right)^2 \right] \left( \frac{N}{N-1} \right) \quad (\text{A12-4})$$

Now  $P = \sum X/N$ , so Eq. (A12-4) may be written for proportions.

$$\sigma_X^2 = (P - P^2) \left( \frac{N}{N-1} \right) = PQ \left( \frac{N}{N-1} \right) \quad (\text{A12-5})$$

Now, remembering that when  $X$  can take only the values of 0 and 1,  $p = \bar{X}$ , we substitute the value  $\sigma_X^2 = PQ \left( \frac{N}{N-1} \right)$  from Eq. (A12-5) into Eq. (A12-2), thus obtaining

$$\sigma_p^2 = \frac{PQ}{n} \left( \frac{N}{N-1} \right) \left( 1 - \frac{n}{N} \right)$$

or

$$\sigma_p^2 = \frac{PQ}{n} \left( \frac{N-n}{N-1} \right)$$

Also since

$$\sigma_d^2 = n^2 \sigma_p^2$$

for infinite populations, it follows that

$$\sigma_d^2 = nPQ \left( \frac{N-n}{N-1} \right)$$

for finite populations.



## Some Elements of Bayesian Decision Theory

Up to this point we have followed classical statistical methods in the formulation of a decision rule in that we have selected a value of  $\alpha$  on the basis of intuition. When determining an optimal sample size, as in the appendix to Chapter 10, we specified points on a power curve without an explicit statement of the reasons why these points were chosen. Although the considerations leading to selection of certain values of  $\alpha$  and  $\beta$  were not explicitly given, they were implicitly recognized.<sup>(1)</sup> It will be recalled that in Sec. 10.5,  $\alpha$  was said to depend upon both the degree of belief in the null hypothesis and the cost associated with rejecting a true null hypothesis. Bayesian methods, however, make explicit use of both prior probabilities (degrees of belief) and estimated losses (or costs) resulting from various decisions, as well as costs of sampling and measurement.

The illustrations in this chapter will come from the field of statistical quality control. The reader is warned that these illustrations are greatly oversimplified in order to allow concentration on the fundamental principles and to reduce the volume and level of mathematics involved. This chapter may be omitted on first reading without destroying the continuity of presentation.

---

<sup>(1)</sup> If there were no criteria for determining the sample size, acceptance of a two-sided hypothesis would mean only that the sample size was too small to require  $H_0$  to be rejected. If  $n$  is sufficiently large, the null hypothesis for almost any two-sided test will be rejected.

### 13.1 TWO-SIDED TESTS

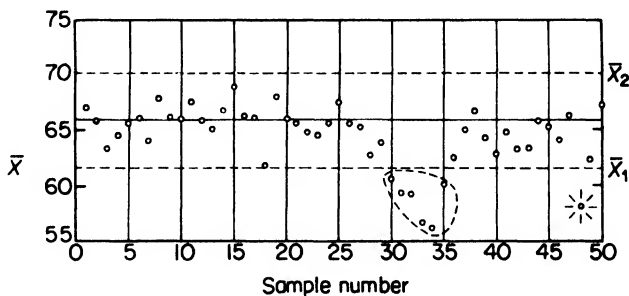
Let us illustrate a two-sided test using a control chart for means. Before doing this, we shall discuss the rationale for the control chart method.

**Control Charts.** As the name indicates, one of the purposes of statistical quality control is to keep the manufacturing process under control. When a manufacturing process is governed by a constant system of causes operating in a random manner, it is said to be *in control*. Thus if all non-random types of variability have been eliminated or taken into consideration quantitatively and if we have discovered the probability distribution of the random fluctuations, the process is in control. If the process is in control we can state the probability that an item taken at random will fall within specified limits. Thus the essence of control is *predictability*.

A state of controlled quality is desirable for at least two reasons. (1) We can determine whether the quality of the manufactured product is satisfactory. The quality is satisfactory if we can predict that almost all of the items will be within specified limits, i.e., if the natural tolerances are inside the engineering tolerances. A process may be in control and yet may produce too many defective items, either because the average value is too large or too small, or because there is too much variability in the process. (2) We have a sound basis for making specifications. There is no point in making specifications so tight that they cannot be enforced economically. On the other hand, if the natural tolerances are far inside the engineering tolerances, it may be appropriate to change the latter.

The characteristic tool of process control is the *control chart*, which is a graphic method of presenting a sequence of samples. Chart 13.1 is typical. On this chart are shown the mean warp-breaking strengths ( $\bar{X}$ -values) of

CHART 13.1: CONTROL CHART FOR MEANS: WARP-BREAKING STRENGTH IN POUNDS OF 50 SAMPLES OF 4 ITEMS EACH OF COTTON CLOTH



Source: Dudley J. Cowden and William S. Connor, "The Use of Statistical Methods for Economic Control in Industry," *Southern Economic Journal*, Vol. XII, (October 1945), pp. 115-29.

successive samples of 4 units each of cotton cloth. The samples were taken at intervals of approximately one hour. It will be noticed that the means fluctuate around a central line and for the most part inside two broken lines known as *control limits*. The lower control limit is indicated symbolically as  $\bar{X}_1$  and the upper control limit, as  $\bar{X}_2$ . The central line is an estimate of the standard value of the process mean, often the mean of many previous sample means. Whenever a point goes outside the control limits (samples 30 through 35 and 48 in this case), trouble is indicated; the foreman is immediately notified to look for the source of the trouble. The trouble will, of course, be corrected if it is found.

The control limits are located statistically in such a way that the process engineer will usually look for serious trouble but will not waste his time looking for nonexistent trouble. The control limits are supposed to strike an economical balance between two kinds of errors: (1) looking for trouble that does not exist, and (2) failing to look for trouble that does exist. The probability of committing either of these kinds of error should not be unduly large, yet neither should be reduced to such an extent that it unduly increases the other.

The control chart is a valuable tool because: (1) it gives early warnings of trouble; (2) it is flexible. Referring to Chart 13.1, we see that sample 30 is the first to go outside the control limits. Yet an alert person might suspect impending trouble earlier by noticing the downward trend in the plotted points. Another type of warning may be too long a run of items above (or below) the central line. In fact, any unusual arrangement of plotted points may cause suspicion, in some cases strong enough to lead to action.

Chart 13.1 is for controlling the average value of the product. We want assurance that the mean value of the output varies only from random causes, but we also want assurance that the variability of the process changes only from random causes. Therefore, a control chart for means is often accompanied by a control chart for ranges. Because of the computational labor, control charts for standard deviations are unusual.

All control charts are similar in that they are characterized by a *central line*  $\theta_0$  and by *control limits*. In general we may say that the control limits are

$$\theta_0 \pm \delta \sigma_{\theta} \quad (13-1)$$

where  $\theta_0$  is the standard value of the parameter,  $\sigma_{\theta}$  is the standard error of the sampling distribution of the statistic, and  $\delta$  is a constant. Typically,  $\delta = 3$ .

For a control chart for arithmetic means, the control limits are

$$\mu_0 \pm 3\sigma_{\bar{x}}$$

For a control chart for ranges we have<sup>(1)</sup>

$$R_0 \pm 3\sigma_R$$

<sup>(1)</sup> Generally  $\sigma$  is estimated by computing  $a_0\bar{R}$ , where  $\bar{R}$  is the historical average value of the ranges. The lower,  $LCL(R)$ , and upper,  $UCL(R)$ , control limits for ranges are  $D_3\bar{R}$  and  $D_4\bar{R}$ , respectively. See Appendix 9 for values of  $a_0$ ,  $D_3$ , and  $D_4$ .

Often a control chart for proportion defective is used. In some cases such a chart is a substitute for a control chart for means. The chart is easy to interpret and easy to form, since it is usually less difficult to say that an item is good or bad than to say *how* good or bad it is. Because proportions are not distributed normally for small samples, a control chart for proportions typically uses sample sizes of 100 or more when  $\delta = 3$ . If  $P_0$  is less than 0.01,  $n$  should be 200 or more. The control limits for  $p$  charts are usually

$$P_0 \pm 3\sigma_p$$

Also, instead of a control chart for proportion defective, one for number of defectives  $d$  is often used.

Finally, a control chart for number of defects  $c$  is sometimes used, the control limits being

$$c_0 \pm 3\sigma_c$$

Regardless of the kind of control chart, there are three interrelated statistical problems: (1) the optimum sample size  $n$ , (2) the optimum number of standard errors from the central line to a control limit  $\delta$ , (3) the frequency of sampling. We shall now address ourselves to the first two of these problems with reference to the arithmetic mean.

**Optimal Control Chart for Means.** Let us assume that the process mean  $\mu$  can take on various values during different periods of time, though at the time a sample is taken it can assume only one value. We also assume that we have subjective estimates of the probability of different values of  $\mu$ . Let us call these subjective probabilities *prior probabilities*. These prior probabilities,  $\text{Prob}(\mu)$ , are given in Table 13.1, and we notice that they sum to unity. We also notice that the prior probability distribution is approximately normal<sup>(3)</sup> with standard deviation  $\sigma$  of 12.28 mm. Table 13.1 also shows the losses that will be incurred by accepting or rejecting the process for each stated value of  $\mu$ . Accepting or rejecting the process is, of course, equivalent to accepting or rejecting the null hypothesis that the process is in control. The losses stated in this table should not be thought of as losses from operation, but merely as reductions in gain. The table shows that no loss is involved if we accept the process when  $\mu = 100$ . This value of  $\mu$  is the value of the central line on the control chart (or more generally the value of  $\mu$  when the null hypothesis is true). As usual, we will denote this mean as  $\mu_0$ .

Notice that, as  $\mu$  departs from its standard value,  $\mu_0$ , the loss involved in accepting the process increases because of the deteriorated quality of the output. On the other hand, the loss involved in rejecting the process is a maximum when the mean is at its standard level. This loss may be thought

<sup>(3)</sup> This prior probability distribution is an oversimplification. Actually  $\mu$  can take on any value within some range instead of only seven values. Another simplification is that we do not consider how often the samples should be taken. In a similar way the cost functions are oversimplified in that we assume that monetary costs are an adequate representation of "utilities," an assumption which may or may not be true (see Problem 1).

TABLE 13.1: PRIOR PROBABILITIES AND LOSSES

$\mu$	Prior Prob ( $\mu$ )	CONDITIONAL LOSS IF	
		Accept $L_A$	Reject $L_R$
70	0.016	\$45	\$0
80	0.094	30	10
90	0.234	20	15
100	0.312	0	25
110	0.234	20	15
120	0.094	30	10
130	0.016	45	0
Total	1.000	...	...

of as the cost of looking for nonexistent trouble, or the loss resulting from making unneeded adjustments in the process. The farther the process mean departs from its standard value, the easier it is to find the trouble, and hence the smaller the loss involved in rejecting the process. (In the general case the losses resulting from making various decisions will arise from different causes.)

In addition to the information provided by Table 13.1, we must have two additional pieces of information. First, we must know the standard deviation of the  $X$  values, which is considered to be independent of the process mean and is estimated with very small error to be  $\sigma = 10$  mm.<sup>(4)</sup> Second, we must know the cost of sampling, which is estimated to be 5¢ per unit sampled.

Before we undertake to determine the "best" decision rule, let us look at an extreme case. Suppose that we choose not to have a control chart at all and thus eliminate the cost of sampling. Suppose also that we arbitrarily accept the process. The expected loss under this rule will be

$$\Sigma [\text{Prob}(\mu) \cdot L_A] = (0.016)45 + \cdots + (0.016)45 = \$16.44$$

Notice that this figure is simply the expected value of the probability distribution associated with  $L_A$ , i.e., with the various values of  $\mu$  corresponding to the various values of  $L_A$ . On the other hand, if we decide to reject the process arbitrarily, the expected loss is

$$\Sigma [\text{Prob}(\mu) \cdot L_R] = (0.016)0 + \cdots + (0.016)0 = \$16.70$$

Of these two rules, it is obviously better to accept the process at all times, since the expected loss is smaller, although only marginally. Our next question is, "Can we reduce these loss figures by sampling?" If we can, the control chart is justified. Also, the logic of the problem requires that we go still further and ask: "What values of  $\delta$  and  $n$  will minimize our expected loss?"

<sup>(4)</sup> This value is determined from past experience rather than from current sampling.

As a guess, let us try the following decision rule: Take a random sample of size  $n = 16$  and set control limits at  $\mu_0 \pm 2\sigma_{\bar{x}}$ ; reject the process if the sample mean falls outside of the control limits. The basis of the decision rule can be described more concisely as  $n = 16$ ,  $\delta = 2$ . In terms of classical statistical methods, to set control limits two standard deviations from the hypothetical population mean is the same as setting alpha at 0.04550, since 4.550 percent of the area under the normal curve lies beyond  $\pm 2$  standard deviations from the mean.

Under this sampling plan the lower control limit is

$$\begin{aligned}\bar{X}_1 &= \mu_0 - 2\sigma_{\bar{x}} \\ &= 100 - 2(2.5) = 95\end{aligned}$$

and the upper control limit is

$$\begin{aligned}\bar{X}_2 &= \mu_0 + 2\sigma_{\bar{x}} \\ &= 100 + 2(2.5) = 105\end{aligned}$$

since

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{4} = 2.5$$

Table 13.2 illustrates the conditional probabilities of accepting the process, given various values of  $\mu$ ,  $\text{Prob}(\text{Accept} \mid \mu)$ . The table also shows the conditional probabilities of rejecting the hypothesis, given various values of  $\mu$ ,  $\text{Prob}(\text{Reject} \mid \mu)$ . Of course, we recall that accepting the process is the same as accepting the null hypothesis,  $\mu = \mu_0$ . Notice the similarities between this table and Table 10.1. Also notice that the column  $\text{Prob}(\text{Reject} \mid \mu)$  is the power function of this test,  $1 - \beta(\mu)$ . [The column  $\text{Prob}(\text{Accept} \mid \mu)$  is the operating characteristic function of the test,  $\beta(\mu)$ ].

Table 13.3 utilizes the information given in Tables 13.1 and 13.2 for the purpose of computing average expected loss. Only the  $\bar{L}$  column requires explanation. It is the average loss, the two conditional probabilities being used as weights, and is given by

$$\bar{L} = \text{Prob}(\text{Accept} \mid \mu) \cdot L_A + \text{Prob}(\text{Reject} \mid \mu) \cdot L_R$$

Then, the *expected* average loss is the sum of the entries in the last column

$$E(\bar{L}) = \sum [\text{Prob}(\mu) \cdot \bar{L}]$$

and is \$9.31. To this figure we add the sampling cost of  $16(\$0.05) = \$0.80$  to get a *total* expected loss of \$10.11. This final figure represents the loss consequence of the sampling plan and is much lower than either of our previously determined "no sample" plans

Tables 13.1 through 13.3 involve a great deal of duplication that can be eliminated once the procedure is understood. The material has been presented in small bites to aid the digestion.

TABLE 13.1: COMPUTATION OF PROBABILITY OF ACCEPTING AND REJECTING THE PROCESS FOR DIFFERENT VALUES OF  $\mu$   
 ( $n = 16$ ,  $\delta = 2$ ,  $\mu_0 = 100$ ,  $\sigma\bar{x} = 2.5$ ,  $\bar{X}_1 = 95$ ,  $\bar{X}_2 = 105$ )

$\mu$	$\bar{X}_1 - \mu$	$\bar{X}_2 - \mu$	$z_1$	$z_2$	$P(z_1)$	$Q(z_2)$	$Prob (Accept   \mu)^*$	$Prob (Reject   \mu)^\dagger$
70	25	35	10	14	1.00000	0.00000	0.00000	1.00000
80	15	25	6	10	1.00000	0.00000	0.00000	1.00000
90	5	15	2	6	0.97725	0.00000	0.02275	0.97725
100 = $\mu_0$	-5	5	-2	2	0.02275	0.02275	0.95450	0.04550 = $\alpha$
110	-15	-5	-6	-2	0.00000	0.97725	0.02275	0.97725
120	-25	-15	-10	-6	0.00000	1.00000	0.00000	1.00000
130	-35	-25	-14	-10	0.00000	1.00000	0.00000	1.00000

\*  $Prob (Accept | \mu) = 1.0 - Prob (Reject | \mu) = \beta(\mu)$

†  $Prob (Reject | \mu) = P(z_1) + Q(z_2) = 1 - \beta(\mu)$

See Table 10.1 and Appendix 1.

**TABLE 13.3: COMPUTATIONS FOR DETERMINATION OF EXPECTED AVERAGE LOSS**

$\mu$	$L_A$	$L_R$	$Prob (Accept   \mu)$	$Prob (Reject   \mu)$	$\bar{L}$	$Prob (\mu)$	$Prob (\mu) \cdot \bar{L}$
70	\$45	\$0	0.00000	1.00000	\$0.000	0.016	\$0.000
80	30	10	0.00000	1.00000	10.000	0.094	0.940
90	20	15	0.02275	0.97725	15.114	0.234	3.537
100	0	25	0.95450	0.04550	1.138	0.312	0.355
110	20	15	0.02275	0.97725	15.114	0.234	3.537
120	30	10	0.00000	1.00000	10.000	0.094	0.940
130	45	0	0.00000	1.00000	0.000	0.016	0.000

Expected average loss. . . . . \$9.31

Let us now, following the method of Table 13.3, try other values of  $\delta$ , holding  $n$  constant at 16. The results are as follows:

$\delta$	<i>Total expected loss</i>
3.0	\$10.09
2.5	9.95
2.0	10.11

We tentatively conclude that the optimum value of  $\delta$  is close to 2.5. Holding  $\delta$  constant at 2.5, we try different values of  $n$ :

$n$	<i>Total expected loss</i>
25	\$10.26
16	9.95
9	10.17

It seems likely that the best combination is in the neighborhood of  $\delta = 2.5$ ,  $n = 16$ . Let us tabulate some further results:

$\delta$	$n$	9	16	25
3.0		\$10.55	10.09	\$10.23
2.5		10.17	9.95	10.26
2.0		10.08	10.11	10.51

These results further confirm our previous findings. If it is desirable to make estimates of other values of  $\delta$  and  $n$  within the range of the previous table, we may continue by trial and error in the same way we have illustrated. Other methods of attack are possible but will not be illustrated here.<sup>(6)</sup>

<sup>(6)</sup> For example, we may fit response surfaces by the method of steepest ascent or descent. See Acheson J. Duncan, *Quality Control and Industrial Statistics*, 3rd ed. (Homewood, Ill.: Richard D. Irwin, 1965), Chapter XXXVII.



A further saving of time and effort can be obtained by using *opportunity loss* instead of conditional loss. Opportunity loss (sometimes called *regret*) is the *extra* loss incurred by not making the correct decision for each value of the parameter. Table 13.4 illustrates the computation of expected

**TABLE 13.4: COMPUTATION OF EXPECTED AVERAGE OPPORTUNITY LOSS**  
( $n = 16, \delta = 2$ )

$\mu$	Prob (Wrong decision)	$\mathcal{L}$	$\bar{\mathcal{L}}$	Prob ( $\mu$ )	Prob ( $\mu$ ) $\cdot \mathcal{L}$
70	0.00000	\$45	\$0.000	0.016	\$0.000
80	0.00000	20	0.000	0.094	0.000
90	0.02275	5	0.114	0.234	0.027
100 = $\mu_0$	0.04550	25	1.138	0.312	0.355
110	0.02275	5	0.114	0.234	0.027
120	0.00000	20	0.000	0.094	0.000
130	0.00000	45	0.000	0.016	0.000
Expected opportunity loss . . . . .					\$.041

opportunity loss for the decision rule  $n = 16, \delta = 2$ . The loss values of Table 13.4 are  $\mathcal{L} = L_A - L_R$  for each value of  $\mu$  except where  $\mu = \mu_0$ . In this case  $\mathcal{L} = L_R - L_A$ . (If there were more than two possible acts, the calculation would be modified in an obvious way.) The probability of a wrong decision, Prob (Wrong decision) is Prob (Accept |  $\mu$ ) in every case except where  $\mu = \mu_0$ , in which case it is Prob (Reject |  $\mu$ ). Notice that the expected average opportunity loss is \$0.41, which is \$8.90 less than the expected average loss obtained in Table 13.3. It can be shown that this difference of \$8.90 will remain constant for each possible decision rule so that the same optimum rule will be obtained by either method (see Problem 2).

### 13.2 ONE-SIDED TESTS

To illustrate the application of Bayesian inference to one-sided tests, we shall use the subject of product control. A synonym for product control is acceptance sampling.

**Acceptance Sampling.** The purpose of acceptance sampling is to decide whether to accept or reject a lot on the basis of evidence afforded by one or more samples drawn at random from the lot in question. If the lot is always either accepted or rejected on the basis of one sample, we have *single sampling*. Sometimes a relatively small sample is taken, and if the quality of the sample is very good or very bad, the lot is then either accepted or rejected. Otherwise, a second sample is taken, and the lot is accepted or rejected on the basis of the two samples combined. This procedure is known as

*double sampling.* *Multiple sampling* is a logical extension of double sampling. However, the decision to accept or reject need not be made after the first or second sample is taken, but it may be made after drawing several samples. We will confine ourselves in this section to single sampling plans.

**Optimum Single Sampling Plan.** Suppose that a manufacturer receives a lot of four items, which he will process and later market. The profit that he can hope to make depends in large part upon the quality of the incoming lots. Upon receipt of a lot he is faced with the decision of whether to accept or reject it. His decision will be made upon the basis of a single sample.

From long experience the manufacturer is convinced that over the long run the average number of defective items  $D$  in a lot is 2. The manufacturer also thinks that it is reasonable to suppose that  $D$  is distributed according to the binomial probability distribution.<sup>(6)</sup> Thus, the manufacturer estimates that the expected value of  $D$  is  $E(D) = 2$ . Using the binomial distribution, he can estimate the "long run" fraction defective  $P$  in the population to be

$$\hat{P} = \frac{E(D)}{N} = \frac{2}{4} = 0.5$$

and since

$$\hat{Q} = 1 - \hat{P} = 0.5$$

the binomial distribution

$$\text{Prob}(D) = \binom{N}{D} \hat{P}^D \hat{Q}^{N-D}$$

can be used to estimate the prior probability distribution of  $D$  as given in Table 13.5.

TABLE 13.5: PRIOR PROBABILITIES AND LOSSES

$D$	Prior Prob ( $D$ )*	CONDITIONAL LOSS IF	
		Accept $L_A$	Reject $L_R$
0	0.0625	\$0.00	\$3.00
1	0.2500	1.00	3.00
2	0.3750	2.00	3.00
3	0.2500	4.00	3.00
4	0.0625	6.00	3.00

\*  $\binom{4}{D} 0.50^4$ .

<sup>(6)</sup> Using  $N = 4$  permits us to consider each possible population parameter. This consideration could not be achieved in the last section. However, there is a loss of realism in that a lot size this small would be very rare. Finally, it should be noted that the parameter  $D$  need not be distributed according to the binomial distribution, or any known distribution for that matter. We only assert that this distribution is thought to be the appropriate one in this case.

Table 13.5 gives as well  $L_A$  and  $L_R$  as discussed in the last section. The loss function  $L_A$  indicates that losses resulting from acceptance of a lot of high defective content are greater than for lots of low defective content. Also, delays in production and extra processing cause these losses to increase more rapidly than the lot quality deteriorates. The loss resulting from rejecting the lot is considered to be constant at \$3.00. Shipping costs associated with returning the lots to the sender might behave in this manner. The problem is simplified by assuming that the manufacturer can always obtain a sufficient number of lots for current production.

Again, let us define a "no sample" plan. If the lot is accepted arbitrarily, the expected loss is

$$\Sigma [\text{Prob}(D) \cdot L_A] = (0.0625)0.00 + \dots + (0.0625)6.00 = \$2.38$$

and if the lot is rejected arbitrarily, the expected loss is

$$\Sigma [\text{Prob}(D) \cdot L_R] = (0.0625)3.00 + \dots + (0.0625)3.00 = \$3.00$$

Of the two rules it is better to accept all lots.

Now let us define another plan. Under this plan let us take a random sample, without replacement, of size  $n = 2$  and accept the lot if we find one or fewer defective items. This plan can be described concisely as  $n = 2$ ,  $A = 1$ , where  $A$  is the acceptance number. The rejection number  $R$  is  $A + 1$ .

As in the last section we wish to find  $\text{Prob}(\text{Accept} | D)$  and  $\text{Prob}(\text{Reject} | D)$ , where the words "accept" and "reject" refer to accepting and rejecting the lot. If the lot is accepted, the number of defective items in the sample  $d$  must be less than or equal to one. Thus

$$\text{Prob}(\text{Accept} | D) = \text{Prob}(d < 1 | D)$$

and, as in the last section

$$\begin{aligned} \text{Prob}(\text{Reject} | D) &= 1 - \text{Prob}(\text{Accept} | D) \\ &= \text{Prob}(d > 2 | D) \end{aligned}$$

Table 13.6 shows the calculation of  $\text{Prob}(d = 0 | D)$  for the five values

**TABLE 13.6: CALCULATION OF  $\text{Prob}(d = 0 | D)$ , USING HYPERGEOMETRIC DISTRIBUTION  $N = 4$ ,  $n = 2$**

$D$	$d$	$\binom{D}{d}$	$G$	$g$	$\binom{G}{g}$	$\binom{D}{d} \binom{G}{g}$	$\text{Prob}(d = 0   D)^*$
0	0	1	4	2	6	6	1.0
1	0	1	3	2	3	3	0.5
2	0	1	2	2	1	1	0.166...
3	0	1	1	2	0	0	0.0
4	0	1	0	2	0	0	0.0

$$^* \binom{N}{n} = \binom{4}{2} = 6.$$

of  $D$ . The hypergeometric distribution is used in this case, since the sample size ( $n = 2$ ) is large relative to the population size ( $N = 4$ ). We recall from Sec. 12.3 that

$$\text{Prob}(d = 0 | D) = \frac{\binom{D}{d} \binom{G}{g}}{\binom{N}{n}}$$

where  $G = N - D$  and  $g = n - d$ . Table 13.7 calculates in the same way

**TABLE 13.7: CALCULATION OF  $\text{Prob}(d = 1 | D)$ , USING HYPERGEOMETRIC DISTRIBUTION  $N = 4$ ,  $n = 2$**

$D$	$d$	$\binom{D}{d}$	$G$	$g$	$\binom{G}{g}$	$\binom{D}{d} \binom{G}{g}$	$\text{Prob}(d = 1   D)^*$
0	1	0	4	1	4	0	0.0
1	1	1	3	1	3	3	0.5
2	1	2	2	1	2	4	0.666...
3	1	3	1	1	1	3	0.5
4	1	4	0	1	0	0	0.0

$$* \binom{N}{n} = \binom{4}{2} = 6.$$

$\text{Prob}(d = 1 | D)$  for the various values of  $D$ . Then

$$\text{Prob}(\text{Accept} | D) = \text{Prob}(d = 0 | D) + \text{Prob}(d = 1 | D)$$

and is shown in Table 13.8 along with  $\text{Prob}(\text{Reject} | D)$ .

In the same way as in the last section, we proceed to calculate in Table 13.8

**TABLE 13.8: COMPUTATIONS FOR DETERMINATION OF EXPECTED AVERAGE LOSS**

$D$	$L_A$	$L_R$	$\text{Prob}(\text{Accept}   D)$	$\text{Prob}(\text{Reject}   D)$	$\bar{L}$	$\text{Prob}(D)$	$\text{Prob}(D) \cdot \bar{L}$
0	\$0.00	\$3.00	1.0000	0.0000	\$0.000	0.0625	\$0.000
1	1.00	3.00	1.0000	0.0000	1.000	0.2500	0.250
2	2.00	3.00	0.8333	0.1667	2.167	0.3750	0.813
3	4.00	3.00	0.5000	0.5000	3.500	0.2500	0.875
4	6.00	3.00	0.0000	1.0000	3.000	0.0625	0.187

Expected average loss . . . . . \$2.12

the expected average loss

$$E(\bar{L}) = \sum [\text{Prob}(D) \cdot \bar{L}]$$

which is \$2.12. Assuming that sampling costs are 5¢ per item sampled, we find that this sampling rule is superior to either of the “no sample” rules, since it has associated with it a total expected loss of \$2.22.

Again, the decision rule just discussed may or may not be the best one available. To find out, we would need to investigate all combinations of sample sizes and acceptance numbers. Excluding the "no sample" plans, we see that there will be 14 such plans, since  $A$  cannot exceed  $n$ . The determination of the optimal plan will be left as an exercise (see Problem 3).

**Relationship to Bayes' Theorem.** Let us now investigate the consequences of the rule  $n = 2$ ,  $A = 1$  using a different computational technique. This technique is more laborious, but it has the advantage of making explicit the relationship between the methods used in this chapter and Bayes' theorem.

From Sec. 6.11 we recall that we may write the conditional probability

$$\text{Prob}(D \mid d < 1) = \frac{\text{Prob}(D \cap d < 1)}{\text{Prob}(d < 1)}$$

$$\text{or} \quad \text{Prob}(D \mid \text{Accept}) = \frac{\text{Prob}(D) \cdot \text{Prob}(\text{Accept} \mid D)}{\text{Prob}(\text{Accept})} \quad (13-2)$$

Also, it follows that

$$\text{Prob}(D \mid \text{Reject}) = \frac{\text{Prob}(D) \cdot \text{Prob}(\text{Reject} \mid D)}{\text{Prob}(\text{Reject})} \quad (13-3)$$

The probabilities given by Eqs. (13-2) and (13-3) are called *posterior* probabilities. They represent a modification of the prior probabilities according to the results of sampling.

Table 13.9 calculates  $\text{Prob}(D \mid \text{Accept})$  and finds  $\sum [\text{Prob}(D \mid \text{Accept}) \cdot L_A]$ . Table 13.10 carries out the same calculation for  $\text{Prob}(D \mid \text{Reject})$ . The third column in Table 13.9 is the same as the fourth column in Table 13.8. The fourth column in Table 13.9 is the product of the second and third columns. Notice that the sum of the fourth column entries, which is 0.7500, is the unconditional probability,  $\text{Prob}(\text{Accept})$ . The fifth column in Table 13.9 is found by dividing each entry in the fourth column by  $\text{Prob}(\text{Accept}) = 0.7500$ . Thus the fifth column is  $\text{Prob}(D \mid \text{Accept})$  as given by Eq. (13-2). Table 13.10 proceeds in the same way, using Eq. (13-3).

Now if we multiply each of these expected loss figures, respectively, by the unconditional probabilities that the lot will be accepted or rejected, we have

$$\begin{aligned} & \sum [\text{Prob}(D \mid \text{Accept}) \cdot L_A] \cdot \text{Prob}(\text{Accept}) \\ & \quad + \sum [\text{Prob}(D \mid \text{Reject}) \cdot L_R] \cdot \text{Prob}(\text{Reject}) \end{aligned}$$

$$\text{or} \quad 1.83(0.7500) + 3.00(0.2500) = \$2.12$$

This figure is the same expected average loss figure that we have previously established (see Table 13.8). The student is advised to carry out this procedure, using the illustration of the last section.

TABLE 13.9: CALCULATION OF  $\Sigma [Prob(D | Accept) \cdot L_A]$ 

$D$	Prior Prob ( $D$ )	Conditional Prob (Accept   $D$ )	Joint Prob (Accept $\cap D$ )	Posterior Prob ( $D$   Accept)	$L_A$	Prob ( $D$   Accept) $\cdot L_A$
0	0.0625	1.0000	0.0625	0.0833	\$0.00	\$0.000
1	0.2500	1.0000	0.2500	0.3333	1.00	0.333
2	0.3750	0.8333	0.3125	0.4167	2.00	0.833
3	0.2500	0.5000	0.1250	0.1667	4.00	0.667
4	0.0625	0.0000	0.0000	0.0000	6.00	0.000
Total	1.0000	...	0.7500	1.0000	...	\$1.83

TABLE 13.10: CALCULATION OF  $\Sigma [Prob(D | Reject) \cdot L_R]$ 

$D$	Prior Prob ( $D$ )	Conditional Prob (Reject   $D$ )	Joint Prob (Reject $\cap D$ )	Posterior Prob ( $D$   Reject)	$L_R$	Prob ( $D$   Reject) $\cdot L_R$
0	0.0625	0.0000	0.0000	0.0000	\$3.00	\$0.000
1	0.2500	0.0000	0.0000	0.0000	3.00	0.000
2	0.3750	0.1667	0.0625	0.2500	3.00	0.750
3	0.2500	0.5000	0.1250	0.5000	3.00	1.500
4	0.0625	1.0000	0.0625	0.2500	3.00	0.750
Total	1.0000	...	0.2500	1.0000	...	\$3.00

**Relationship to Classical Methods.** In terms of the classical methods of hypothesis testing presented in the previous chapters, we have used a sample size  $n = 2$ , an acceptance number  $A = 1$ , and a rejection number  $R = A + 1 = 2$ . A corresponding classical hypothesis is

$$H_0: D = 2$$

$$H_1: D > 2$$

We conduct the test using the hypergeometric distribution and  $\alpha = 0.167$ . The probability of committing an error of the first kind,  $\alpha$ , was determined after the formulation of the decision rule rather than before. It is

$$\alpha = \text{Prob}(d > R \mid D = R)$$

Using the techniques of Sec. 12.3, let us verify these statements.

*Hypotheses:*

$$H_0: D = 2$$

$$H_1: D > 2$$

*Criterion of Significance:*

$$\alpha = 0.167$$

*Rejection Number:*

Table 13.11, which is calculated in the same manner as Table 12.4,

**TABLE 13.11: HYPERGEOMETRIC DISTRIBUTION**

$$N = 4, n = 2, D = 2$$

$d$	$g$	$\text{Prob}(d)$	$\text{Cumulative Prob}(d), Q(d)$
0	2	0.167	1.000
1 = $A$	1	0.667	0.833
2 = $R$	0	0.167	0.167 = $\alpha$

tabulates the hypergeometric distribution for this problem. The probability of obtaining two or more defectives is 0.166... with  $N = 4$ ,  $D = 2$ , and  $n = 2$ . Hence  $A = 1$  and  $R = 2$ . The student should be able to derive the power curve for this test with ease.

It has been said that Bayesian statistical inference differs from classical inference in that it asks, "What is the probability of the hypothesis given the sample?" whereas classical inference asks, "What is the probability of the sample given the hypothesis?" Bayesian inference explicitly treats the parameter as if it were a random variable. Implicitly, classical inference treats the parameter as if it were a random variable in the advice given on the setting of alpha. We recall that alpha is set in classical inference according to two

major criteria: (1) The greater the degree of belief in the null hypothesis (prior probability), the smaller should be the value of alpha. (2) The greater the costs associated with an error of the first kind, the smaller should be the value of alpha. In short, it may be said that Bayesian statistical inference seeks to specify more exactly the two criteria needed to establish the value of alpha by explicit inclusion of prior probabilities and costs in the analysis.

## PROBLEMS

1. St. Petersburg Paradox. A person offers to pay you  $2^k$  dollars where  $k$  is the number of times a fair coin is tossed by you, in a row, before tails comes up. Find the expected monetary value of this game. Would you be willing, or even able, to pay the expected monetary value of this game for the privilege of playing it? Comment on the implications your findings might have for using monetary costs in business decisions.

2. Using the problem of Sec. 13.1, show that the plan  $\delta = 2.5$ ,  $n = 25$  gives expected opportunity loss of \$8.90 smaller than its expected average loss.

3. Using the problem of Sec. 13.2, complete the following table and hence find the optimum sampling plan.

**TOTAL EXPECTED LOSS**

$n \backslash A$	0	1	2	3	4
1			...	...	...
2		\$2.22		...	...
3					...
4					



# 14

## Simple Linear Regression

It is frequently very useful to be able to explain the variations in one series by comparing them with variations in a related series. For instance, a concern buys large quantities of materials that must be of specified tensile strength; these materials must be tested, but the only way of measuring tensile strength involves stretching a unit until it breaks, thereby destroying its usefulness. However, it appears that materials that possess great tensile strength are also likely to be hard, and those that are unusually weak tend to be extremely soft. If the association between hardness and tensile strength is sufficiently close, it is possible to estimate with considerable accuracy the tensile strength of any piece on the basis of the relationship between hardness and tensile strength as shown by a sample. This is an economical method of testing, since the test for hardness is not destructive.

### 14.1 THE SCATTER DIAGRAM

In the second and third columns of Table 14.1 are shown the hardness  $X_2$  and tensile strength  $X_1$  of a random sample of 27 pieces of wrought aluminum alloy, arranged according to hardness as shown by tests. This, of course, is not the order in which the pieces were tested. The arrangement is used to enable one to see that, as hardness increases, there is a tendency for tensile strength to increase. We would like to know if tensile strength  $X_1$  is associated in a linear fashion with hardness  $X_2$  and if it can be predicted from hardness.

Although Table 14.1 enables us to observe each item separately, we cannot gain a clear impression of the nature of the relationship

**TABLE 14.1: COMPUTATION OF VALUES USED IN DETERMINING MEASURES OF RELATIONSHIP BETWEEN HARDNESS AND TENSILE STRENGTH OF 27 PIECES OF WROUGHT ALUMINUM ALLOY (HARDNESS IS MEASURED IN UNITS OF BRINNELL HARDNESS; TENSILE STRENGTH REPRESENTS THOUSANDS OF POUNDS PER SQUARE INCH.)**

Rank in hardness (lowest to highest)	Hardness $X_2$	Tensile strength $X_1$	$X_2^2$	$X_2X_1$	$X_1^2$
1	16	8	256	128	64
2	24	14	576	336	196
3	26	15	676	390	225
4	27	13	729	351	169
5	28	16	784	448	256
6	29	16	841	464	256
7	30	16	900	480	256
8	35	19	1,225	665	361
9	41	23	1,681	943	529
10	42	26	1,764	1,092	676
11	44	25	1,936	1,100	625
12	45	26	2,025	1,170	676
13	49	30	2,401	1,470	900
14	52	30	2,704	1,560	900
15	52	30	2,704	1,560	900
16	59	37	3,481	2,183	1,369
17	64	33	4,096	2,112	1,089
18	70	38	4,900	2,660	1,444
19	80	39	6,400	3,120	1,521
20	87	52	7,569	4,524	2,704
21	95	48	9,025	4,560	2,304
22	99	60	9,801	5,940	3,600
23	99	61	9,801	6,039	3,721
24	101	59	10,201	5,959	3,481
25	116	68	13,456	7,888	4,624
26	119	71	14,161	8,449	5,041
27	120	67	14,400	8,040	4,489
Total:	1,649	940	128,493	73,631	42,376
Mean:	61.07407	34.81481	...	...	...
Correction term:	...	...	*100,711.1	*57,409.6	*32,725.9
Variation or covariation:			27,781.9	16,221.4	9,650.1

Source: Aluminum Research Laboratories, Aluminum Company of America.

$X_1$  and  $X_2$  are carried to 7 digits to permit formal internal checks. Actually, only 3 or 4 digits are significant.

\* These correction terms, which are subtracted from the sum of squares and cross products in order to obtain measures of variation and covariation, are

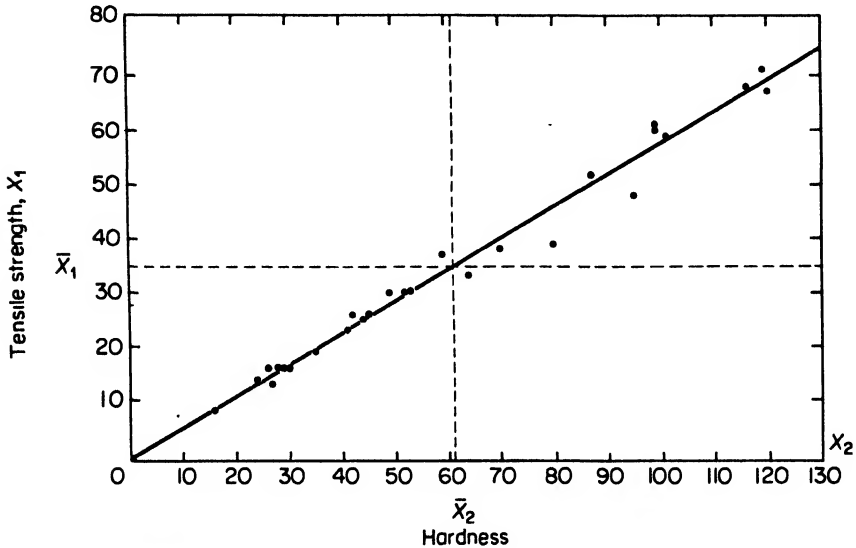
$$\frac{(\sum X_2)^2}{n} = \bar{X}_2 \sum X_2; \quad \frac{(\sum X_1)(\sum X_2)}{n} = \bar{X}_2 \sum X_1 = \bar{X}_1 \sum X_2; \quad \frac{(\sum X_1)^2}{n} = \bar{X}_1 \sum X_1.$$

between the two variables; if hardness increases a given number of points, we do not know how much tensile strength increases. Also, we do not know whether the relationship between the two variables is linear or if it behaves in some more complicated manner. To aid in the answering of

these questions we plot the sample observations, by placing dots on coordinate paper, one dot for each pair of observations. This *scatter diagram*, which depicts the scatter of our sample observations, is shown in Chart 14.1 and leaves us with the visual impression that the two variables are rather closely related in a linear fashion.

The straight line running through the scatter diagram is the “best” fitting line that can be drawn, according to one criterion of goodness-of-fit. The

**CHART 14.1: SCATTER DIAGRAM AND SAMPLE REGRESSION LINE FOR HARDNESS AND TENSILE STRENGTH OF 27 PIECES OF WROUGHT ALUMINUM ALLOY.**



Source: Table 14.1.

equation is of the type

$$\hat{X}_1 = a + bX_2 \quad (14-1)$$

and the symbol  $\hat{X}_1$  denotes the estimated value of  $X_1$  for a given value of  $X_2$ . Equation (14-1) is known as a *linear estimating equation* or *sample linear regression equation*. As we shall see, the line drawn through the scatter of points in Chart 14.1 has the equation

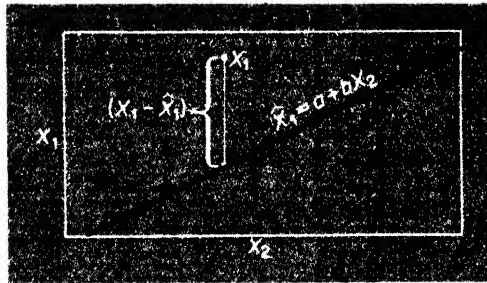
$$\hat{X}_1 = -0.84534 + 0.58388X_2$$

and the line is called the *estimating* or *sample regression line*. Once determined, the estimating equation is useful in estimating the value of  $X_1$ . For example, when  $X_2 = 27$  the estimate is

$$\hat{X}_1 = -0.84534 + 0.58388(27) = 14.9$$

## 14.2 THE LEAST SQUARES CRITERION

It is evident from an inspection of Chart 14.1 that the estimating line does *not* pass through every one of the points on the scatter diagram. Indeed, as we shall see, it need not pass through any of them. Since we are interested in estimating tensile strength  $X_1$  on the basis of hardness, it might be logical to view the values  $X_1 - \hat{X}_1$  as being *errors of the estimate* (see the diagram below). It would seem natural to want the sum of these errors, taken without



regard to sign, to be as small as possible. The least squares criterion demands that the estimating line be fitted to the scatter of points in such a way that the sum of the squares of the vertical deviations about the estimating line be as small as or smaller than for any other straight line fitted to the scatter of points. Thus, the least squares criterion demands that

$$\sum (X_1 - \hat{X}_1)^2 \quad (14-2)$$

be a minimum.<sup>(1)</sup>

Notice carefully that we have actually established a theoretical model of the relationship between the variables  $X_1$  and  $X_2$  in concentrating on the vertical deviations of  $X_1$  about  $\hat{X}_1$ . We have postulated that the variable  $X_2$  is *not* a random variable but is given without error. Any errors in the estimation of  $X_1$ , when the regression equation is used, are attributable to deviations in the variable  $X_1$  about the regression line. These errors may arise because even though the true value of  $X_1$ , given  $X_2$ , lies directly on the regression line, errors in the measurement of the true value of  $X_1$  obscure this value. The deviations might also arise because there is no true value of  $X_1$  for a given value of  $X_2$ . There may in fact be infinitely many such values clustered about a mean value. To put it another way, there may be a basically unpredictable element in  $X_1$  that cannot be explained by  $X_2$ . In our example, tensile strength may also vary with temperature, humidity, and so on.

<sup>(1)</sup> Other models are logical under certain conditions. For example, we may wish to minimize the horizontal deviations about  $\hat{X}_1$  or the perpendicular deviations about  $\hat{X}_1$ .

### 14.3 THE NORMAL EQUATIONS AND THEIR SOLUTION

Performing the mathematics necessary to minimize expression (14-2) leads to the formation of two simultaneous equations that are often called *normal equations*.<sup>(2)</sup> The equations are

$$\left. \begin{array}{ll} na + b \sum X_2 = \sum X_1 & \text{I} \\ a \sum X_2 + b \sum X_2^2 = \sum X_1 X_2 & \text{II} \end{array} \right\} \quad (14-3)$$

Using the data given in Table 14.1, we see that for our problem

$$\begin{array}{ll} 27a + 1649b = 940 & \text{I} \\ 1649a + 128,493b = 73,631 & \text{II} \end{array}$$

These equations may be solved by using elementary algebra for  $a = -0.84534$  and  $b = 0.58388$ , which are the intercept and slope of the straight line plotted in Chart 14.1.<sup>(3)</sup>

Solution of the normal equations is facilitated by noting that when I is divided by  $n$

$$a = \bar{X}_1 - b\bar{X}_2 \quad (14-4)$$

Also, it is easy to show that<sup>(4)</sup>

$$b = \frac{\sum x_1 x_2}{\sum x_2^2} \quad (14-5)$$

We already know that the terms  $\sum x_1^2$  and  $\sum x_2^2$  are referred to as *variation*

<sup>(2)</sup> The word "normal" does not imply any connection with the normal probability distribution but relates to the mathematical form of the equations. The student with a knowledge of intermediate differential calculus can derive these equations without much trouble. Let  $e = X_1 - \hat{X}_1$  and set the two partial derivatives  $\partial \sum e^2 / \partial a$  and  $\partial \sum e^2 / \partial b$  equal to zero, which is a necessary condition for a minimum.

<sup>(3)</sup> Multiply I by  $\bar{X}_2 = 61.07407$  and subtract the result from II to obtain  $b$ .

$$\begin{array}{rcl} 1649a + 128,493b = 73,631 & \text{II} & \\ 1649a + 100,711b = 57,410 & \text{I} \times (61.07407) & \\ \hline 27,782b = 16,221 & & \\ b = 0.5839 & & \end{array}$$

Then, substitute the value of  $b$  into either I or II to obtain  $a$ . The calculations may be checked by direct insertion into Eqs. (14-3).

<sup>(4)</sup> Multiply I by  $\bar{X}_2 = \sum X_2 / n$  and subtract the result from II to give

$$b[\sum X_1^2 - (\sum X_2)^2 / n] = \sum X_1 X_2 - (\sum X_1 \sum X_2) / n$$

Then, by Eqs. (14-6) and (14-7),  $b = \sum x_1 x_2 / \sum x_2^2$ . An alternative way of expressing the slope which is sometimes useful is  $b = \sum x_2 X_1 / \sum x_1^2$  (see Problem 4).

and may be calculated as follows

$$\Sigma x^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{n} \quad (14-6)$$

Similarly, *covariation* is found by use of

$$\Sigma x_1x_2 = \Sigma X_1X_2 - \frac{(\Sigma X_1)(\Sigma X_2)}{n} \quad (14-7)$$

The expression  $(\Sigma X)^2/n$  is called the *correction term* for variation, and the expression  $(\Sigma X_1)(\Sigma X_2)/n$  is called the correction term for covariation.

Computation of the basic elements needed to find the slope and intercept of the estimating equation as well as other statistics needed in this chapter may be conveniently expressed as follows:

Matrix of sums of squares and cross products	Matrix of — correction terms	Matrix of = variation and covariation
----------------------------------------------------	------------------------------------	---------------------------------------------

Or, placing numbers across the top and left sides of the matrices to aid the memory, we have

$$\begin{array}{cc} (1) & (2) \end{array} \quad \begin{array}{cc} (1) & (2) \end{array} \quad \begin{array}{cc} (1) & (2) \end{array}$$

$$\begin{array}{l} (1) \left[ \Sigma X_1^2 \quad \Sigma X_1X_2 \right] \\ (2) \left[ \Sigma X_2X_1 \quad \Sigma X_2^2 \right] \end{array} - \begin{array}{l} (1) \left[ \frac{(\Sigma X_1)^2}{n} \quad \frac{\Sigma X_1 \Sigma X_2}{n} \right] \\ (2) \left[ \frac{\Sigma X_2 \Sigma X_1}{n} \quad \frac{(\Sigma X_2)^2}{n} \right] \end{array} = \begin{array}{l} (1) \left[ \Sigma x_1^2 \quad \Sigma x_1x_2 \right] \\ (2) \left[ \Sigma x_2x_1 \quad \Sigma x_2^2 \right] \end{array}$$

Using the data in Table 14.1 and omitting the redundant southwest term, we have

$$\begin{bmatrix} 42,376 & 73,631 \\ & 128,493 \end{bmatrix} - \begin{bmatrix} 32,725.9 & 57,409.6 \\ & 100,711.1 \end{bmatrix} = \begin{bmatrix} 9650.1 & 16,221.4 \\ & 27,781.9 \end{bmatrix}$$

Then

$$b = \frac{\Sigma x_1x_2}{\Sigma x_2^2} = \frac{16,221.4}{27,781.9} = 0.5838837$$

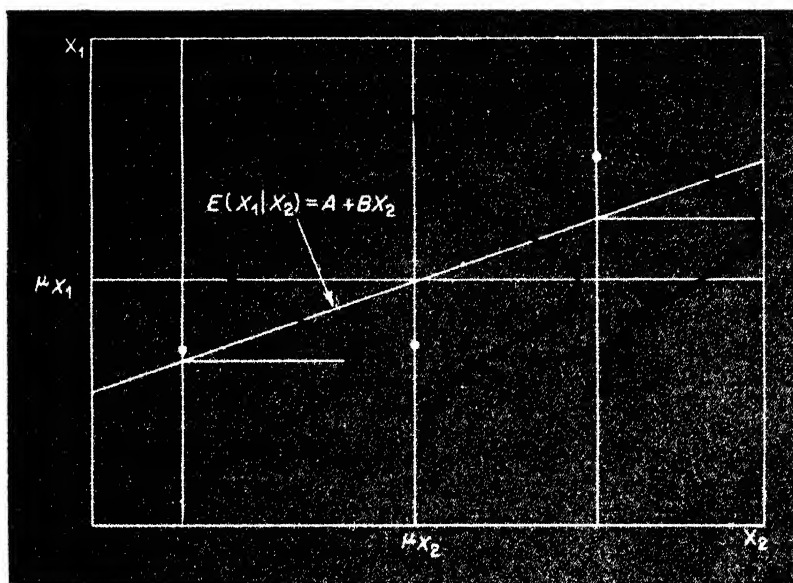
We have carried this calculation to excessive decimal places to allow for accurate calculation of  $a$ , which is

$$a = \bar{X}_1 - b\bar{X}_2 = 34.81481 - 0.5838837(61.07407) = -0.84534$$

#### 14.4 RELATIONSHIP OF THE SAMPLE REGRESSION EQUATION TO THE POPULATION

The sample regression equation computed in the last section was based upon a set of sample values for  $X_1$  and  $X_2$ . It is, therefore, an estimate of a population regression equation. In order to test hypotheses and set confidence limits for the population parameters, we must make some

**CHART 14.2: POPULATION REGRESSION LINE AND CONDITIONAL DISTRIBUTIONS OF  $X_1 | X_2$ .**



assumptions about the probability distribution of the elements in the population. In Chart 14.2 a hypothetical population regression line with intercept  $A$  and slope  $B$  is shown. We assume that we may pick a specific value of  $X_2$  and visualize a distribution of values for  $X_1$  in the population which is *conditional* upon the chosen value of  $X_2$ . To simplify the discussion, we assume further that each of these distributions of  $X_1$  is normal and that each has the same variance. Also, we postulate that the conditional mean of each of these distributions is given by the population regression line.<sup>(5)</sup> That is to say

$$E(X_1 | X_2) = A + BX_2 \quad (14-8)$$

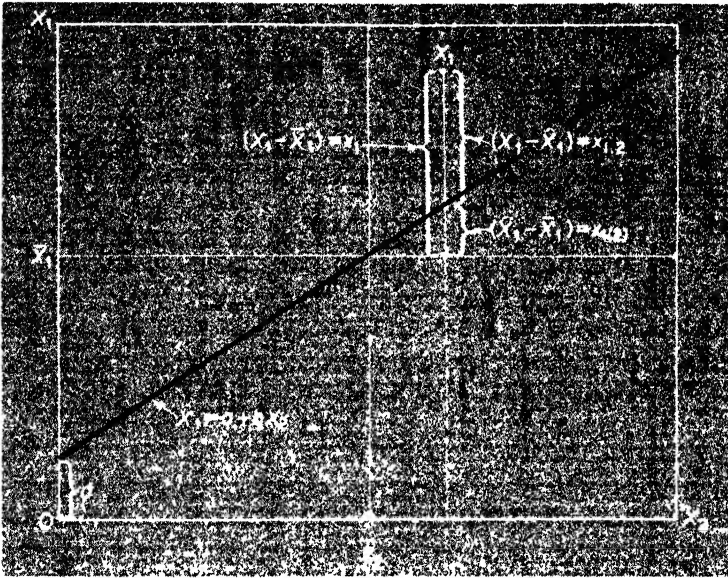
The variance of each conditional distribution is defined in the same manner as that given in previous sections for infinite populations.

$$\sigma^2_{X_1|X_2} = E[X_1 | X_2 - E(X_1 | X_2)]^2 \quad (14-9)$$

The scatter diagram referred to previously represents sample pairs of observations taken from these conditional distributions. Three such sample pairs

<sup>(5)</sup> These assumptions are "classical" ones for "normal regression." Under these conditions the estimates of  $A$  and  $B$  derived by least squares will be identical to those which could be derived by the method of maximum likelihood. For an excellent discussion of the consequences of altering these assumptions see: J. Johnston, *Econometric Methods*, (New York: McGraw-Hill Book Company, 1960).

**CHART 14.3: PARTITIONING THE TOTAL DEVIATION OF A HYPOTHETICAL VALUE OF  $X_1$  INTO COMPONENTS WHICH HAVE BEEN EXPLAINED AND NOT EXPLAINED BY THE REGRESSION LINE.**



are indicated by the dots in Chart 14.2. It is obvious that we cannot take all of the samples from the same conditional distribution if we wish to estimate the population regression equation; thus at least two values of  $X_2$  must be chosen. If the sampling technique is a random one, the sample points will tend to cluster in some random manner about the population regression line. Notice carefully that none of the sample points need lie directly upon the population regression line. Therefore, since the sample regression line will be calculated from these sample points, it need not be identical to the population regression line. However, it can be shown that  $a$  and  $b$  (the sample regression line intercept and slope) are unbiased minimum variance estimators of  $A$  and  $B$  (the population regression line intercept and slope) under the postulated conditions. Thus

$$E(a) = A \quad (14-10)$$

$$E(b) = B$$

Just as we may estimate  $A$  and  $B$  from a sample, so we may estimate  $\sigma^2_{X_1|X_2}$  from a sample. We do this by partitioning the variation in the sample values of  $X_1$  into two parts: a part "explained" by the sample regression line and a part "unexplained" by the sample regression line.

In Chart 14-3 is shown a hypothetical sample value of the variable  $X_1$  associated with a given value of  $X_2$ . Also shown is a hypothetical sample regression line. Notice first that the sample regression line passes through the



point  $(\bar{X}_2, \bar{X}_1)$ , the unconditional sample means of the two variables. This is, in fact, the *only* point on the scatter diagram through which the sample regression line must pass. The fact that the sample regression line must pass through this point is easily shown by using Eq. (14-4).

$$\bar{X}_1 = a + b\bar{X}_2$$

In a similar manner the population regression line must pass through  $\mu_{X_1}$  and  $\mu_{X_2}$ , as is shown in Chart 14.2.

The total deviation of the value of  $X_1$  about  $\bar{X}_1$  is given by

$$X_1 - \bar{X}_1 = \text{Total deviation} \quad (14-11)$$

which is simply the distance between  $X_1$  and the unconditional sample mean of  $X_1$ . For the sample point shown in Chart 14.3, however, the distance between  $X_1$  and the sample regression line is *less* than the distance between  $X_1$  and  $\bar{X}_1$ . In this instance we can say that part of the deviation of  $X_1$  from  $\bar{X}_1$  has been “explained” by use of the sample regression line. Thus

$$\hat{X}_1 - \bar{X}_1 = \text{Explained deviation} \quad (14-12)$$

Notice carefully that the word “explained” is not to be taken in a causal sense; it refers only to reduction in total deviation. Finally, there remains a component of deviation that is not explained by use of the sample regression line.

$$X_1 - \hat{X}_1 = \text{Unexplained or residual deviation} \quad (14-13)$$

It should be clear that for the given value of  $X_1$ , these components of deviation are additive.

$$\begin{array}{ccccc} (X_1 - \bar{X}_1) & = & (\hat{X}_1 - \bar{X}_1) & + & (X_1 - \hat{X}_1) \\ \text{Total} & & \text{Explained} & & \text{Unexplained} \\ \text{deviation} & & \text{deviation} & & \text{deviation} \end{array}$$

For concise notation we may express these components of deviation as

$$x_1 = x_{1(2)} + x_{1.2} \quad (14-14)$$

The parentheses will be used arbitrarily to represent the word “explained” and the dot will be used to represent the word “unexplained.” Summing the squares of these deviations, we obtain components of variation which are also additive.

$$\begin{array}{ccccc} \Sigma x_1^2 & = & \Sigma x_{1(2)}^2 & + & \Sigma x_{1.2}^2 \\ \text{Total} & & \text{Explained} & & \text{Unexplained} \\ \text{variation} & & \text{variation} & & \text{variation} \end{array} \quad (14-15)$$

Table 14.2 illustrates one method of calculating these measures of variation and shows that they are additive. A more practical method of calculating these measures is given in Table 14.3. Notice in particular that

$$\Sigma x_{1(2)}^2 = b \Sigma x_1 x_2$$

**TABLE 14.2: COMPUTATIONS ILLUSTRATING CONCEPTS OF THE THREE TYPES OF DEVIATIONS AND THREE TYPES OF VARIATION: UNEXPLAINED; EXPLAINED; TOTAL. (UNEXPLAINED, EXPLAINED AND TOTAL DEVIATIONS ARE COMPUTED FROM TABLE 14.1.)**

Item number	$X_1$	$\hat{X}_1$	$x_{1.2} = X_1 - \hat{X}_1$	$x_{1(2)} = \hat{X}_1 - \bar{X}_1$	$x_1 = X_1 - \bar{X}_1$	$x_{1.2}^2$	$x_{1(2)}^2$	$x_1^2$
1	8	8.4970	-0.4970	-26.3178	-26.8148	0.25	692.63	719.03
2	14	13.1680	0.8320	-21.6468	-20.8148	0.69	468.58	433.26
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
26	71	68.6366	2.3634	33.8218	36.1852	5.59	1143.91	1309.37
27	67	69.2205	-2.2205	34.4057	32.1852	4.93	1183.75	1035.89
Total	940	940.0005	-0.0005	0.0009	0.0004	178.8	9471.3	9650.1*

\* Because of rounding error, there is a discrepancy of one unit in the last digit between total and the sum of unexplained and explained variation. Also notice that, except for rounding error,  $\sum X_1 = \sum \hat{X}_1$ .

It should be obvious that, other things equal, the larger the proportion of the total variation in the sample values of  $X_1$  which has been explained by the use of the sample regression line, and the smaller the proportion unexplained, the more reliable is the sample regression line. This is an important concept and will be used extensively in later chapters.

Using the components of sample variation, we may now compute three measures of sample variance. First,

$$s_1^2 = \frac{\sum x_1^2}{n-1} \quad (14-16)$$

is familiar to us as the unbiased estimator of the unconditional population

**TABLE 14.3: CALCULATION OF COMPONENTS OF VARIATION**

Source of variation	Computational formulas	Example values
Total: $\sum x_1^2$	$\sum X_1^2 - \frac{(\sum X_1)^2}{n}$	9650.1 (see Table 14.1)
Explained: * $\sum x_{1(2)}^2$	$b \sum x_1 x_2$	0.58388(16,221.4) = 9471.3
Unexplained: $\sum x_{1.2}^2$	$\sum x_1^2 - \sum x_{1(2)}^2$	9650.1 - 9471.3 = 178.8

\*  $\sum (\hat{X}_1 - \bar{X}_1)^2 = \sum (a + bX_2 - \bar{X}_1)^2$ . But  $a = \bar{X}_1 - b\bar{X}_2$ , so  $\sum (\hat{X}_1 - \bar{X}_1)^2 = \sum (\bar{X}_1 - b\bar{X}_2 + bX_2 - \bar{X}_1)^2 = \sum [b(X_2 - \bar{X}_2)]^2$  or  $\sum (\hat{X}_1 - \bar{X}_1)^2 = b^2 \sum x_2^2 = b \sum x_1 x_2$ .

variance of the variable  $X_1$ . Second,<sup>(6)</sup>

$$s_{1.2}^2 = \frac{\sum x_{1.2}^2}{n - 2} \quad (14-17)$$

is, under the assumptions set out earlier, an unbiased estimator of  $\sigma_{X_1|X_2}^2$ . The number of degrees of freedom for this estimator is  $n - 2$ , since two restrictions are placed on the sample values of  $X_1$  by  $\sum x_{1.2}^2$ , namely, the slope and intercept of the sample regression equation. Intuitively,  $s_{1.2}^2$  estimates  $\sigma_{X_1|X_2}^2$ , because just as  $\sigma_{X_1|X_2}^2$  measures the variance of the population values of  $X_1$  about the population regression line, so  $s_{1.2}^2$  measures the variance of the sample values of  $X_1$  about the sample regression line.

Finally, since variation is additive, it would seem reasonable for the number of degrees of freedom for the estimates of variance to be additive and

$$s_{1(2)}^2 = \frac{\sum x_{1(2)}^2}{1} \quad (14-18)$$

is the sample variance associated with explained variation. The divisor in Eq. (14-18) is 1 because of the original  $n - 1$  degrees of freedom associated with  $\sum x_1^2$ ;  $n - 2$  degrees of freedom were associated with  $\sum x_{1.2}^2$ , leaving one degree of freedom to be associated with  $\sum x_{1(2)}^2$ .

## 14.5 TESTS OF HYPOTHESES AND CONFIDENCE LIMITS FOR THE SLOPE AND INTERCEPT

In order to test hypotheses and establish confidence limits for  $A$  and  $B$  we must, as in the past, find the variances for  $a$  and  $b$  and define appropriate sampling distributions for these statistics. To do this, we define the estimates of the variances of  $a$  and  $b$ .

$$s_a^2 = \frac{s_{1.2}^2 \sum X_2^2}{n \sum x_2^2} \quad (14-19)$$

$$s_b^2 = \frac{s_{1.2}^2}{\sum x_2^2} \quad (14-20)$$

The estimated standard errors of  $a$  and  $b$  are found by evaluating the square root of their respective estimated variances. From the data in Table 14.3

$$s_{1.2}^2 = \frac{\sum x_{1.2}^2}{n - 2} = \frac{178.8}{27 - 2} = 7.15$$

<sup>(6)</sup> Notice that the divisor in Eq. (14-17),  $n - 2$ , may be written  $n - m$ , where  $m$  is the number of constants in the regression equation. In a similar way we may write the divisor for Eq. (14-18) as  $m - 1$ .

and from Table 14.1

$$s_a = \sqrt{\frac{s_{1.2}^2 \sum X_2^2}{n \sum x_2^2}} = \sqrt{\frac{7.15(128,493)}{27(27,781.9)}} = 1.1065$$

$$s_b = \sqrt{\frac{s_{1.2}^2}{\sum x_2^2}} = \sqrt{\frac{7.15}{27,781.9}} = 0.01604$$

It is generally accepted practice to report these standard errors in parentheses under the intercept and slope of the sample regression equation because, as we shall see, they are extremely useful. The reporting procedure, after rounding, is

$$\hat{X}_1 = -0.845 + 0.584X_2$$

(1.106)      (0.016)

The  $t$  distribution with  $n - 2$  degrees of freedom is used in testing hypotheses for both the slope and intercept. This follows because of the fact that  $a$  and  $b$  are both normally distributed with standard errors containing the term  $s_{1.2}^2$ , which is an estimate of  $\sigma_{X_1|X_2}^2$  with  $n - 2$  degrees of freedom. The means of the sampling distributions of  $a$  and  $b$  are  $A$  and  $B$ , respectively.

**Tests of Hypotheses Concerning  $A$ .** In the present problem the estimated value of  $A$  is  $-0.84534$ . Suppose that we are interested in testing

$$H_0: A = 0$$

$$H_1: A \neq 0$$

at  $\alpha = 0.05$ . The hypothetical value of  $A$  is  $A_0 = 0$ . From Appendix 4 we find, in the usual way, the rejection values for  $t$  using  $n - 2 = 25$  degrees of freedom and  $\alpha/2 = 0.025$ . The rejection values are  $\pm 2.060$ . Then the test statistic is formulated in the usual way

$$t = \frac{a - A_0}{s_a} = \frac{-0.84534 - 0.0}{1.1065} = -0.76398$$

and we do not reject the null hypothesis at the stated level of significance;  $t = -0.764 > t_L = -2.060$ . We conclude that the intercept is not significantly different from zero at the stated level of significance.

A one-sided test could be conducted by use of  $t_\alpha$  rather than  $t_{\alpha/2}$ . Also, the hypothetical value of  $A$  need not be zero but may take on any value of interest.

**Tests of Hypotheses Concerning  $B$ .** In the present problem the estimated value of  $B$  is  $0.58388$ . Suppose that we are interested in testing

$$H_0: B = 0$$

$$H_1: B > 0$$

at  $\alpha = 0.05$ . The hypothetical value of  $B$  is  $B_0 = 0$ . From Appendix 4 we find the upper rejection value for  $t$ , which is  $1.708$ , using  $25$  degrees of freedom

and  $\alpha = 0.05$ . Then the test statistic is

$$t = \frac{b - B_0}{s_b} = \frac{0.58388 - 0.0}{0.01604} = 36.4 \quad (14-21)$$

and we reject the null hypothesis at the stated level of significance and conclude that the slope is significantly different from zero. Modifications in the test can be made, as were noted for tests concerning the intercept.

We have stressed tests of significance for the slope and intercept of the regression equation because of their frequent use in business and economic research. For example, in economics it is often desirable to ascertain whether two variables behave proportionately, i.e., to determine whether or not the intercept is significantly different from zero when a two-sided test is used. Also, since the sample regression line must pass through  $\bar{X}_1$ , if its slope is zero, it must be identical to  $\bar{X}_1$ . In this case none of the variation of  $X_1$  about  $\bar{X}_1$  would be explained by the use of the sample regression equation, and the technique of linear regression would be of no analytical value. A look at Appendix 4 will reveal an interesting rule of thumb that is widely used. The rule is that when  $a$  and  $b$  are greater than twice their standard errors, and more than seven pairs of observations are involved, a one-sided test of the hypothesis that  $A$  is zero or that  $B$  is zero will be rejected at  $\alpha = 0.05$ .

**Confidence Limits.** To set  $100(1 - \alpha)$  percent confidence limits for the population intercept, solve

$$a - A_1 = A_2 - a = t_{\alpha/2} s_a$$

and for the population slope solve

$$b - B_1 = B_2 - b = t_{\alpha/2} s_b$$

In both cases the appropriate number of degrees of freedom is  $n - 2$  (see Problem 5).

## 14.6 ESTIMATION

The sample regression equation can be used for two broad purposes. First, it can be used to estimate the *mean* of a conditional distribution of  $X_1$  in the population,  $E(X_1 | X_2)$ , a given value of  $X_2$  being used. Second, it can be used to estimate a *specific value* of  $X_1$  in the population if a specific value of  $X_2$  is given. The estimated standard error of the sample regression line for a fixed value of  $X_2$  is

$$s_{\hat{X}_1} = s_{1.2} \sqrt{\frac{1}{n} + \frac{(X_2 - \bar{X}_2)^2}{\sum x_2^2}} \quad (14-22)$$

To establish  $100(1 - \alpha)$  percent confidence limits for  $E(X_1 | X_2)$ , the value of the population regression line for a specific value of  $X_2$ , we use familiar techniques and solve

$$\hat{X}_1 - E(X_1 | X_2)_1 = E(X_1 | X_2)_2 - \hat{X}_1 - t_{\alpha/2} s_{\hat{X}_1}$$

for a given value of  $X_2$  and use  $n - 2$  degrees of freedom (see Problem 5).

If we calculated and plotted many confidence limits for various values of  $X_2$ , we would have what are called  $100(1 - \alpha)$  percent confidence bands about the sample regression line. These bands would not be linear about the regression line but would become more widely separated from each other the farther the value of  $X_2$  upon which they were based departed from  $\bar{X}_2$ . This means that the reliability of the prediction of the population regression line diminishes as  $X_2$  departs from  $\bar{X}_2$ . The student can verify that Eq. (14-22) is a minimum when  $X_2 = \bar{X}_2$  and, therefore, that the confidence interval is narrowest at this point. This widening of the confidence interval simply reflects that both the slope and intercept of the sample regression equation are subject to error and is forewarning of the hazards of using the sample regression equation to predict the population regression at points remote from  $\bar{X}_2$ .

To set a confidence interval for a specific value of  $X_1$  using a specific value of  $X_2$ , we proceed in the same manner, using the standard error of prediction.

$$s_{\text{pred}} = s_{1,2} \sqrt{1 + \frac{1}{n} + \frac{(X_2 - \bar{X}_2)^2}{\sum x_2^2}} \quad (14-23)$$

This standard error is larger than the one given in Eq. (14-22) and reflects the greater variability in individual population values over that of population averages. Again, the confidence interval increases as  $(X_2 - \bar{X}_2)^2$  increases.

## 14.7 ALTERNATIVE STATEMENTS OF THE SAMPLE REGRESSION EQUATION

The sample regression equation may be written in any of three ways:

$$\hat{X}_1 = a + bX_2$$

$$\hat{X}_1 = \bar{X}_1 + bx_2 \quad (14-24)$$

$$x_{1(2)} = bx_2 \quad (14-25)$$

where  $x_{1(2)} = \hat{X}_1 - \bar{X}_1$  and  $x_2 = X_2 - \bar{X}_2$

The numerical value and meaning of the slope remain the same for any of the alternative forms.

## 14.8 REGRESSION MODELS

A basic regression model was set out earlier in this chapter. Here, the variable  $X_1$  was considered to be distributed in a particular way about fixed values of the variables  $X_2$ . It is common in regression literature to call the variable  $X_1$  the *dependent variable*, the *explained variable*, the *predictand*, or the *regressand*. The variable  $X_2$  is often called the *independent variable*, the *predictor*, or the *regressor*.<sup>(7)</sup>

Sometimes, especially in the social sciences, the variable  $X_2$  cannot be controlled. Fortunately, regression analysis can still be utilized under conditions where the values of  $X_2$  vary from sample to sample, as in sampling from an existing population. However, an associated complication is that it is often impossible to determine which variable is dependent in a causal sense on the other. In fact, it is often postulated that the two variables vary together. Such relationships are termed *bivariate*; there is a probability distribution for the variable  $X_1$  if a specific value of  $X_2$  is given, and there is a probability distribution for the variable  $X_2$  if a specific value of  $X_1$  is given. In the next chapter we will consider an index of association (the correlation coefficient) that will be useful in cases where  $X_1$  and  $X_2$  vary in a bivariate manner.

Another problem in regression analysis concerns measurement error, especially on the independent variable. If the independent variable is subject to measurement error, it will no longer necessarily be true that the sample regression slope will be an unbiased estimator of the population regression slope. The main illustration of this chapter is valid only insofar as we are able to assume that there is little or no measurement error in the reported values of  $X_2$ . This distinction is often neglected in social science presentations of regression analysis but is not to be taken lightly in view of the large measurement errors commonly encountered in social statistics. In certain cases the bias in a regression coefficient may be corrected, but to do so requires estimation of the measurement error. This estimation is usually very difficult to make.

---

## PROBLEMS

1. What is the only point through which the sample regression line must pass? Why is this reasonable?

---

<sup>(7)</sup> Actually, the terms "independent" and "dependent" are vanishing from mathematical literature. However, the terms still seem to be in favor in statistical literature, and we will use them extensively.

2. Draw a scatter diagram of the following data and make a visual estimate of the slope and intercept of the regression line. Be sure to use the information in Problem 1 to aid you in this estimate. Now, use the least squares criterion to fit the line and calculate  $s_a$  and  $s_b$ . Test the significance of the slope and intercept at the 0.05 level (one-sided test).

$X_2$	1	2	3	4	5
$X_1$	2	5	2	4	4

3. Give an algebraic proof of Eqs. (14-24) and (14-25) and verify that the meaning of  $b$  remains the same as for Eq. (14-1).

4. Give an algebraic proof of the following:

a.  $b = \frac{\sum x_2 x_1}{\sum x_2^2}$  (Hint: Expand  $\sum (X_2 - \bar{X}_2)(X_1 - \bar{X}_1)$ ).

b.  $\sum \hat{X}_1 = \sum X_1$ .

5. Using the data in Table 14.1 and other calculations in this chapter that may be helpful, show that:

a. When  $X_2 = 20$  the predicted value of  $X_1$  is approximately 10.83.

b. Ninety-five percent confidence limits for  $E(X_1 | X_2 = 20)$  are approximately 9.11 and 12.55.

c. Ninety-five percent confidence limits for  $B$  are approximately 0.55 and 0.62.

d. Ninety-five percent confidence limits for  $A$  are approximately  $-3.13$  and  $1.44$ .

6. Argue that the conditional variance  $\sigma_{X_1|X_2}^2$  is the same as the "unexplained" population variance  $\sigma_{X_1,2|X_2}^2$  (often called the variance of the *residuals* in the population).

7. Suppose that you wish to fit a regression line without an intercept; i.e.,

$$\hat{X}_1 = bX_2$$

Will the line necessarily pass through the point  $(\bar{X}_2, \bar{X}_1)$ ?



# 15

## The Correlation Coefficient

In the previous chapter we discussed the least squares method of fitting a straight line to a scatter of observations. In this chapter we will discuss the correlation coefficient, which, using the model of Chapter 14, will give us a measure of the closeness of fit of the sample regression line to the scatter of points.

### 15.1 THE STANDARD ERROR OF ESTIMATE

We saw in the last chapter that

$$s_{1,2}^2 = \frac{\sum x_{1,2}^2}{n - 2} \quad (15-1)$$

offers an estimate of the variance displayed by the elements of the population about the population regression line. The square root of this statistic is often called the *standard error of estimate* (or *standard error of regression*).

If  $s_{1,2}^2$  is "small," we would be led to believe that there was little variance of the sample values of  $X_1$  about the sample regression line and, therefore, that the sample regression line offered a close fit to the observed scatter of points. The opposite would be true for a "large" value of  $s_{1,2}^2$ .

One difficulty with using  $s_{1,2}^2$  as a measure of the closeness of fit of the sample regression line to the scatter of points is that it depends upon the units of measurement of the dependent variable. Thus,  $s_{1,2}^2$  is not strictly comparable for different sets of data (see Problem 1). Another difficulty with this measure of closeness of fit is that

it is always positive. What we would like is a measure of closeness of fit that meets these requirements:

1. It does not depend upon the units of measurement of the dependent variable, and, therefore, it is comparable for different sets of data.
2. It gives the sign of the association present, i.e., the sign of the slope of the regression line.

The sample correlation coefficient  $r$  meets these requirements and possesses other useful and interesting statistical properties as well.

## 15.2 TWO ALTERNATIVE CONCEPTS OF THE CORRELATION COEFFICIENT<sup>(1)</sup>

In this section we will show how the correlation coefficient may be thought of as an index of closeness of fit of a sample regression line to a scatter of points under the model of the last chapter. We will also point out how the correlation coefficient may be viewed as a measure of the *covariability* between two variables in a bivariate distribution.

**Square Root of the Proportion of Total Variation that Has Been Explained.** We pointed out in the last chapter that if the variation explained by the sample regression line was large relative to the total variation displayed by the dependent variable, the sample regression line could be considered to offer a close linear fit to the scatter of points. Thus, we define the *coefficient of determination* as the ratio of explained to total variation.<sup>(2)</sup>

$$r^2 = \frac{\sum x_{1(2)}^2}{\sum x_1^2} \quad (15-2)$$

From Table 14.3

$$r^2 = \frac{9471.3}{9650.1} = 0.981$$

Since  $\sum x_{1(2)}^2$  cannot be less than zero or more than  $\sum x_1^2$ , we see that

$$0 < r^2 < 1 \quad (15-3)$$

The coefficient of determination thus always lies in a fixed and convenient range [so long as the denominator of Eq. (15-2) is not zero] and is free from the influence of units of measurement.

<sup>(1)</sup> Other concepts are given in the appendix to this chapter.

<sup>(2)</sup>  $1 - r^2$  is often called the coefficient of nondetermination, whereas  $\sqrt{1 - r^2}$  is often called the coefficient of alienation.

The square root of the coefficient of determination is known as the *coefficient of correlation*. It follows directly from Eq. (15-2) that we may compute this coefficient in the following way:

$$r = \frac{\sum x_1 x_2}{\sqrt{\sum x_1^2 \cdot \sum x_2^2}} \quad (15-4)$$

Since  $\sum x_1 x_2$  determines the sign of the slope of the regression line, the correlation coefficient will assume the same sign as  $b$ .

$$-1 < r < +1 \quad (15-5)$$

A negative correlation coefficient indicates as close a degree of linear association as does a positive correlation coefficient of the same absolute value. When there is no linear association between the sample variables,  $r$  is 0; when the linear association in the sample is perfect, it is  $+1$  or  $-1$ . For the data given in the last chapter

$$r = \frac{16,221.4}{\sqrt{(27,781.9)(9650.1)}} = +0.9907$$

**Geometric Mean of Two Slopes.** We pointed out in Sec. 14.8 that a model in which one variable, the independent variable, is considered fixed while the other variable, the dependent variable, is considered random, is probably less often encountered in the social sciences than a model that postulates that both variables are random. A basic model in the social sciences is, then, one which considers the pairs of sample values ( $X_2, X_1$ ) as being drawn at random from a *bivariate normal population*: i.e., a population where the conditional distribution of ( $X_1 | X_2$ ) is normal and the conditional distribution of ( $X_2 | X_1$ ) is normal. This three-dimensional distribution is sometimes said to resemble a fireman's hat.

Chart 15.1 contains scatter diagrams exhibiting five degrees of correlation. Examination of this chart should aid the student in obtaining an intuitive notion of weak and strong correlation as well as a notion of a bivariate relationship. In addition to the scatter of points given in these diagrams there are two sample regression lines with equations

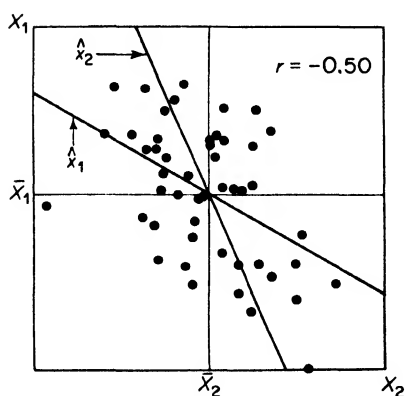
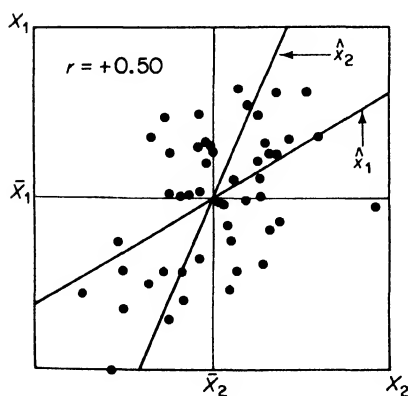
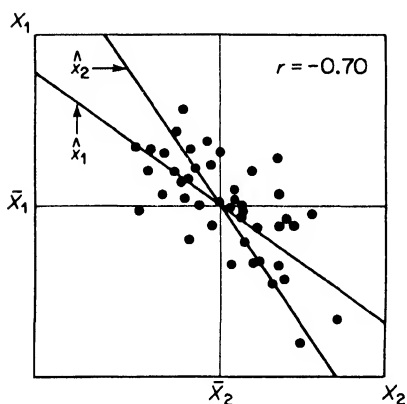
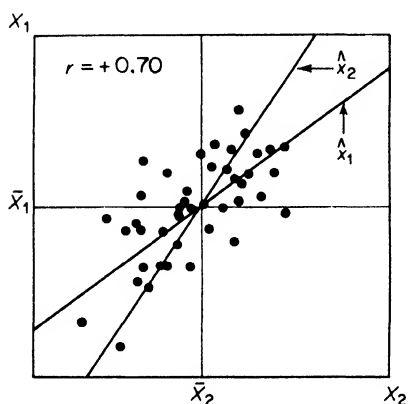
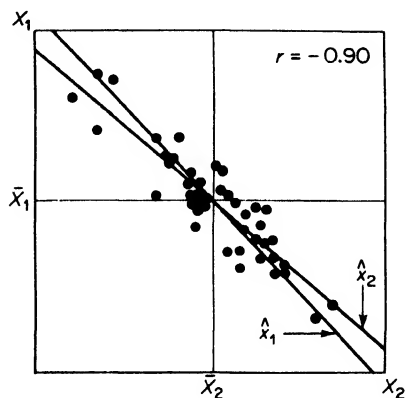
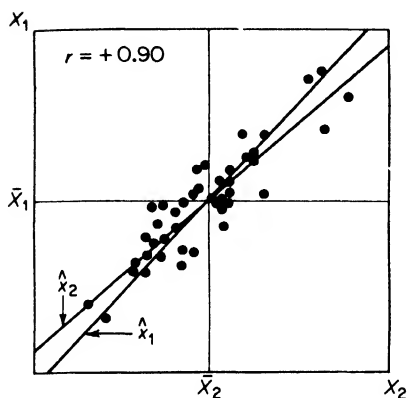
$$\hat{X}_1 = a_{12} + b_{12}X_2 \quad \text{and} \quad \hat{X}_2 = a_{21} + b_{21}X_1$$

the first of which takes  $X_1$  as the dependent variable and the second of which takes  $X_2$  as the dependent variable.<sup>(3)</sup> Unless one is planning to make predictions, there is not necessarily any reason for preferring one equation to the

---

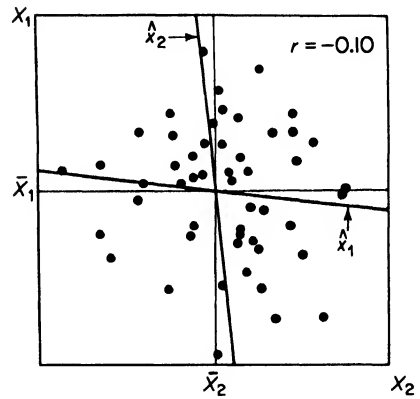
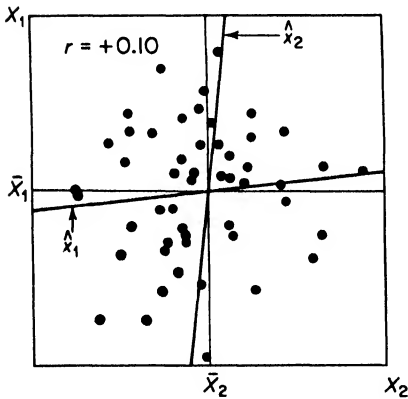
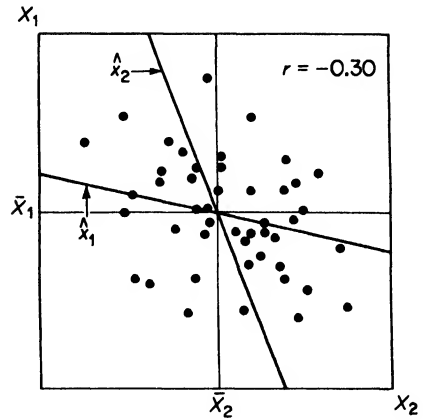
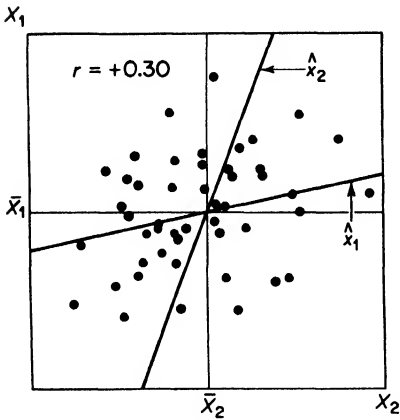
<sup>(3)</sup> The subscripts on the slopes and intercepts are given to distinguish between which variable is being viewed as dependent. The first subscript gives the dependent variable; the second, the independent variable. We will use these subscripts where necessary for clarity of presentation in this and the next chapter.

**CHART 15.1: SCATTER DIAGRAMS SHOWING VARYING DEGREES OF CORRELATION.**



Source: E. C. Fieller, T. Lewis, and E. S. Pearson, *Correlated Random Normal Deviates*, Tracts for Computers No. XXVI (Cambridge: Cambridge University Press, 1955).

CHART 15.1 (cont.).



other. Notice that in the case of strong correlation ( $r = \pm 0.90$ ) the points lie close to the regression lines, whereas in the case of weak correlation ( $r = \pm 0.10$ ) they do not. Also notice that the angle between the two regression lines is smaller in the strong correlation case than it is in the weak correlation case. The correlation coefficient may be thought of as the geometric mean between the slope  $b_{12}$  and the slope  $b_{21}$ . Thus

$$r = \sqrt{b_{12} \cdot b_{21}} \quad (15-6)$$

where

$$b_{12} = \frac{\sum x_1 x_2}{\sum x_2^2} \quad \text{and} \quad b_{21} = \frac{\sum x_2 x_1}{\sum x_1^2}$$

When the two slopes are identical,  $r = 1$ . When one of the slopes is zero,  $r = 0$ . This formula is somewhat deficient because it does not automatically

give the sign of  $r$ , but  $r$  takes the sign of  $b_{12}$  and  $b_{21}$ . The formula is of interest because it is convenient in connection with partial correlation, which will be considered in Chapter 16, and because it stresses the bivariate relationship of many variables encountered in business and economics.

### 15.3 INTERPRETATION OF THE CORRELATION COEFFICIENT

The sample correlation coefficient cannot always be taken at its face value. It is necessary that one know something about the data, and plotting them as a scatter diagram will help greatly in interpreting the measure. Here are a few things to take into consideration.

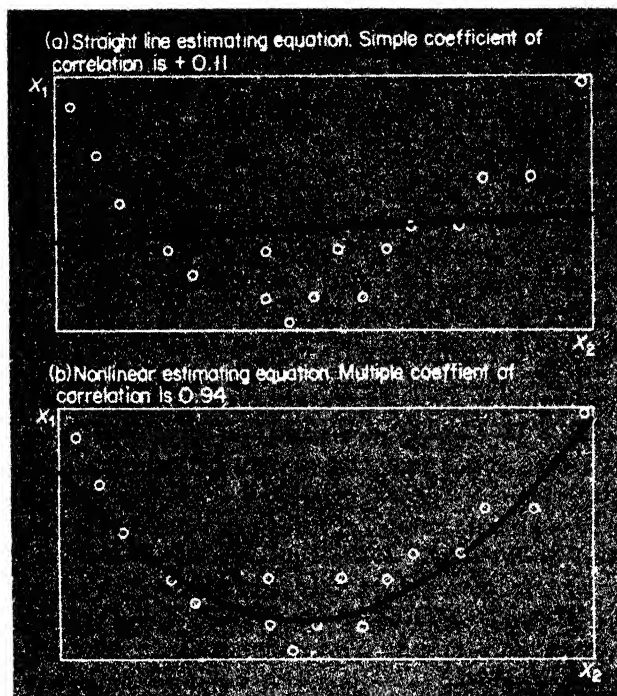
1. *Are the data homogeneous?* In observational data heterogeneity can often be spotted by near bimodality in either the  $X_1$  distribution or the  $X_2$  distribution or by the presence of a few items that are too far out of line with the other items to be considered a matter of chance. If heterogeneity is suspected, different symbols might be used for different sets of data. On the scatter diagram such heterogeneity may show up as a tendency for the dots to cluster into two or more groups, or for one or more dots to be far removed from the others on the chart. Where heterogeneity is observed, it is better to classify the data on some rational basis and correlate each group separately. Individual items clearly governed by a different set of causes should be eliminated before correlating. If these common-sense steps are not taken, one may obtain a misleading impression, not only as to the degree of correlation, but sometimes even as to its sign.

2. *Are the data subject to large errors of measurement?* Since errors in the measurement of the two variables are ordinarily not correlated, such errors reduce the size of  $r$  below its true value. Such *attenuation* can be corrected if the magnitude of the errors is known.<sup>(4)</sup>

3. *Are the data individual measurements, or are they averages?* If the data to be correlated are first grouped into  $k$ -size groups according to the independent variable, if  $\bar{X}_2$  and  $\bar{X}_1$  are computed for each group, and if these means are correlated, the correlation among the means will often be higher than among the individual items taken as a whole (unless  $r = 1$  for the ungrouped data). This effect arises because dispersion of the individual values around the column means has been eliminated. Likewise, if the grouping and averaging are done according to the rows of the dependent variable, the correlation will be increased. Finally, if the data are grouped according to both variables so that there are  $j$  cells, and if  $\bar{X}_2$  and  $\bar{X}_1$  are computed for each cell and these paired cell means correlated, the correlation will be

<sup>(4)</sup> See J. P. Guilford, *Fundamental Statistics in Psychology and Education* (New York: McGraw-Hill Book Co., Inc., 1942), pp. 287-288.

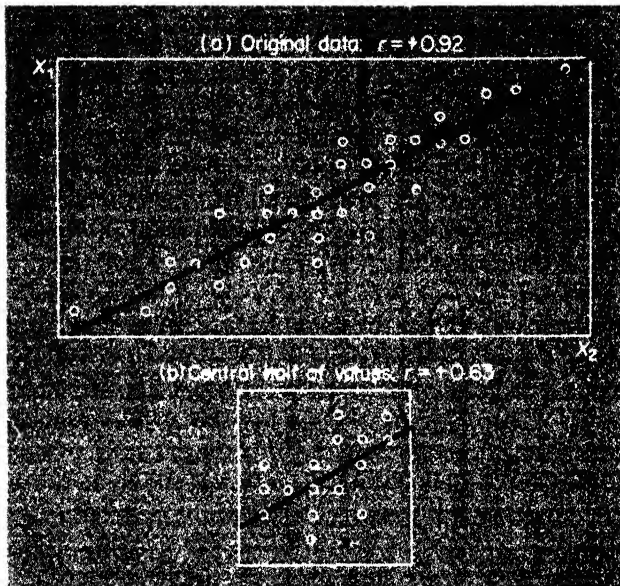
**CHART 15.2: EFFECT ON CORRELATION COEFFICIENT OF USING STRAIGHT LINE ESTIMATING LINE WHEN RELATIONSHIP IS NONLINEAR.**



increased. The increase will be unimportant provided there is a large number of cells.

4. *Is the type of estimating equation appropriate?* If inspection of the scatter diagram reveals that a curved line could more appropriately be fitted to the data than a straight line,  $r$  is a misleading measure of closeness of relationship. A curved line should be fitted, and the coefficient of nonlinear correlation should be computed, the procedures explained in Chapter 16 being followed. So doing will yield a higher coefficient and one that reflects more accurately the closeness of the relationship. For the data of Chart 15.2,  $r = +0.11$ , but by using nonlinear correlation, we find that the correlation is increased to 0.94. Sometimes it may be better to transform one or both of the variables into logarithms, reciprocals, or some other function before correlating (see Problem 6).

5. *Have some of the data been eliminated?* For instance, if retail sales and payrolls are correlated for cities ranging from 100,000 to 500,000 population, the correlation will usually not be so high as if all cities from 10,000 to 5,000,000 are included. This fact is true because retail sales and payrolls are both positively correlated with population, and when the range of values

**CHART 15.3: EFFECT ON CORRELATION COEFFICIENT OF CENSORING DATA**

along both axes is extended,  $\sum x_{1(2)}^2$  is increased without a proportionate increase in  $\sum x_{1,2}^2$ . When the independent variable is controlled, the correlation coefficient is difficult to interpret, and it should not usually be computed.

Consider Chart 15.3. In the scatter diagram at the top all the data are included. The one at the bottom is the same, except that eight large and eight small values of  $X_2$  have been eliminated. The correlation coefficient has been reduced from  $+0.92$  to  $+0.63$ .

6. *Have important variables been neglected?* The simple correlation of earnings of employees with placement scores will usually not be as great as the multiple correlation of earnings with placement scores and experience. Multiple correlation is considered in Chapter 16.

## 15.4 CAUSATION AND THE CORRELATION COEFFICIENT

The coefficient of correlation must be thought of not as something that indicates a particular cause and effect relationship, but only as something that measures degree of association. Any one of the following situations may, in fact, occur.

1. *Either of the variables may be the cause of the other.* The variable that is supposed to be the cause of the variations in the other is usually taken as



the independent variable and plotted along the horizontal axis. The statistician's belief as to which is cause and which is effect is determined by considerations other than the magnitude of  $r$ .

2. *Covariation of the two variables may be due to a common cause or causes affecting each of them in the same way, or in opposite ways.* If it should be found that there is correlation between the number of automobiles registered and the number of telephone subscribers in the various states, it should not be hastily concluded that having an automobile necessitates one's subscribing for telephone service, nor is it easy to see how using a telephone necessitates purchase of an automobile. It is apparent, however, that many persons feel that their incomes are high enough for both items.

3. *The correlation may be due to chance.* Even though there may be no relationship whatever between the variables in the universe from which the sample is drawn, it may be that enough of the paired variables that are selected may vary together, just by accident, to give a fair degree of correlation. Thus, if for a small group of typists we should find positive correlation between foot length and output, we might be inclined to attribute the result to chance. Because we usually deal with samples, and therefore with chance, it is advisable always to test the significance of the correlation coefficient.

## 15.5 TESTS OF SIGNIFICANCE

Reviewing Chart 15.1, we notice that as  $r$  approaches zero, the slope of the regression line given by  $\hat{X}_1 = a_{12} + b_{12}X_2$  approaches zero. It would seem reasonable, therefore, to test the significance of  $r$  by testing the significance of  $b_{12}$ . In fact, as we shall show, testing the significance of  $b_{12}$  is equivalent to testing the significance of  $r$ , and, therefore, both tests need not be performed. However, since there are cases where only  $r$  is known, and because of the statistical interest in the subject, we shall deal with tests of significance for  $r$  rather extensively.

We saw in Sec. 14.5 that a  $t$  statistic, with  $n - 2$  degrees of freedom, could be formulated to test the significance of  $b_{12}$  (test the hypothesis that the hypothetical value of  $B_{12}$  is zero). The statistic is

$$t = \frac{b_{12}}{s_{b_{12}}}$$

and it is easy to show (see Problem 3) that this  $t$  statistic may be written as

$$t = r \sqrt{\frac{n-2}{1-r^2}} \quad (15-7)$$

which follows the  $t$  distribution with  $n - 2$  degrees of freedom.

Using our example data, we have

$$t = 0.9907 \sqrt{\frac{27 - 2}{1 - (0.9907)^2}} = 36.4$$

since  $r = +0.9907$  and  $n = 27$ . Notice carefully that this is the same value for  $t$  that we calculated in Sec. 14.5 using Eq. (14-21).

Let us test the hypothesis that the population correlation coefficient  $\rho$  (rho) is zero, using a one-sided alternative

$$H_0: \rho = 0$$

$$H_1: \rho > 0$$

with  $\alpha = 0.05$ . We enter Appendix 4, and using 25 degrees of freedom and  $\alpha = 0.05$ , we find the upper rejection value for  $t$  to be 1.708. Since the calculated value of  $t$  exceeds the rejection value, we reject the null hypothesis and conclude that  $\rho$  is greater than zero.

The one-sided test was chosen in this case because we are specifically interested in whether or not tensile strength and hardness are positively related. If the sign of the association is not postulated in advance, a two-sided test should be used. The two-sided test differs from the one-sided test only in the fact that  $\alpha/2$ , rather than  $\alpha$ , is used; i.e.,  $\pm t_{\alpha/2}$  are the rejection values for  $t$ .

It is instructive, and it will be especially useful to the student who will study multiple correlation, that Eq. (15-7) may be expressed in terms of the  $F$  distribution when the test concerns a hypothetical population correlation coefficient of zero (see Problem 3).

$$F = \frac{s_{1(2)}^2}{s_{1.2}^2} = \frac{\sum x_{1(2)}^2 / 1}{\sum x_{1.2}^2 / (n - 2)} \quad (15-8)$$

Equation (15-8) is derived by using the fact that  $t = \sqrt{F}$  when the  $F$  ratio has one degree of freedom in the numerator. The table below summarizes the three measures of sample variance whose calculation was begun in Table 14.3.

**TABLE 15.1: ANALYSIS OF VARIANCE FOR TESTING SIGNIFICANCE OF RELATIONSHIP BETWEEN HARDNESS AND TENSILE STRENGTH**

Source of variation	Amount of variation	Degrees of freedom	Estimated variance*
Total	$\sum x_1^2 = 9650.1$	$n - 1 = 26$	$s_1^2 = 371$
Explained	$\sum x_{1(2)}^2 = 9471.3$	1	$s_{1(2)}^2 = 9471$
Unexplained	$\sum x_{1.2}^2 = 178.8$	$n - 2 = 25$	$s_{1.2}^2 = 7.15$

\*  $s_{1(2)}^2$  and  $s_{1.2}^2$  are independent estimates of the population variance under the null hypothesis.

The  $F$  ratio given in Eq. (15-8) emphasizes the fact that if the ratio of explained variance to unexplained variance is "large," given the number of degrees of freedom for the test, the sample correlation coefficient is considered significantly different from zero. The reverse is true if the ratio is "small." This technique of comparing two independent estimates of population variance is known as *analysis of variance* and will be used again in later chapters. Of the three measures of sample variance  $s_{1(2)}^2$  and  $s_{1,2}^2$  are independent of each other but not independent of  $s_1^2$ . Therefore,  $s_1^2$  is not used in the  $F$  ratio.

Using our example data, we see

$$F = \frac{s_{1(2)}^2}{s_{1,2}^2} = \frac{9471}{7.15} = 1325$$

which is the square of the  $t$  statistic computed by Eq. (15-7).

There is a distinction between the  $F$  test and the  $t$  test that is worth noting. The  $F$  table given in this text refers to the probability of obtaining a value of  $r$  as large or larger than that observed in absolute value. Thus for a one-sided test, given the number of degrees of freedom, enter the  $t$  table at  $\alpha$  but the  $F$  table at  $2\alpha$  for comparable rejection values. For a two-sided test enter the  $t$  table at  $\alpha/2$  but the  $F$  table at  $\alpha$ . For example, if  $\alpha = 0.025$ ,  $\nu_1 = 1$ , and  $\nu_2 = 24$ , we find the critical value of  $F$  to be approximately 4.26, using  $Q(F | 1, 24) = 0.05$  for a one-sided test.

Finally, perhaps the easiest method of testing the significance of most simple correlation coefficients is to use a table similar to Appendix 10, where values of  $r_Q$  are given for selected values of  $Q(r | \nu)$ . To test the hypothesis set out earlier in this section, we enter Appendix 10 at  $\nu = n - 2 = 25$  and  $Q(r | \nu) = 0.05$ . The rejection value of  $r$  is given as 0.323, and since our calculated value ( $r = 0.9907$ ) exceeds the rejection value ( $r = 0.323$ ), we reject the null hypothesis. A two-sided test would have used  $Q(r | \nu) = \alpha/2$ . The student can verify that Appendix 10 was constructed by solving Eq. (15-7) for  $r$ , using selected values of  $t_Q$  (see Problem 4).

## 15.6 TESTING OTHER HYPOTHESES AND SETTING CONFIDENCE LIMITS

The  $F$  and  $t$  statistics may be used only to test the hypotheses in which the hypothetical value of  $\rho$  is zero, i.e., to test the significance of  $r$ . The distribution of the correlation coefficient is symmetrical *only* about a population value of zero, in which case it may be approximated by the normal distribution as  $n$  approaches infinity. However, even in the limiting case the distribution is not strictly normal, since the correlation coefficient cannot exceed 1 in absolute value. The tests that follow assume a normal bivariate distribution for the pairs of variables in question.

To test the hypothesis that  $\rho$  is some other value than zero, one may consult David's *Tables*<sup>(5)</sup> or, for an excellent approximation to the exact distribution of the sample correlation coefficient, one may transform  $r$  into  $z_r$ , which is approximately normal in its distribution except when  $n$  is very small.<sup>(6)</sup> Appendix 11 gives values of  $z_r$  for selected values of  $r$ .

Let us now test the following:

*Hypotheses:*

$$H_0: \rho = 0.5$$

$$H_1: \rho > 0.5$$

*Criterion of Significance:*

$$\alpha = 0.10$$

*Rejection Region:*

Using Appendix 3, we obtain  $z_{0.10} = 1.282$

*Standard Error:*

$$\begin{aligned} \sigma_{z_r} &\doteq \sqrt{\frac{3}{3n - 8}} \\ &= \sqrt{\frac{3}{3(27) - 8}} = \sqrt{0.0411} = 0.203 \end{aligned} \quad (15-9)$$

*Test Statistic:*

Interpolating in Appendix 11, we see that when  $r = 0.9907$  and  $\rho = 0.5$ ,

$$z = \frac{z_r - z_\rho}{\sigma_{z_r}} = \frac{2.65 - 0.55}{0.203} = 10.3$$

*Conclusion:*

Since  $z > z_\alpha$ , we reject the null hypothesis and conclude that  $\rho$  is greater than 0.5 at the stated level of significance.

**Confidence Limits.** Confidence limits for  $\rho$  can be set in the familiar manner by solving

$$z_r - z_{\rho_1} = z_{\rho_2} - z_r = z_{\alpha/2} \sigma_{z_r} \quad (15-10)$$

where  $\rho_1$  is the lower confidence limit for  $\rho$  and  $\rho_2$  the upper confidence limit for  $\rho$ .

<sup>(5)</sup> F. N. David, *Tables of the Correlation Coefficient*, Biometrika Office, University College, London, 1938.

<sup>(6)</sup> The  $z$  transformation was formalized by R. A. Fisher. The formula for  $\sigma_{z_r}$  is from Harold Hotelling, "New Light on the Correlation Coefficient and its Transforms," *Journal of the Royal Statistical Society, Series B (Methodological)*, Vol. XV, No. 2, 1953.

Using our data and specifying 90 percent confidence limits, we find that

$$2.65 - z_{\rho_1} = z_{\rho_2} - 2.65 = 1.645(0.203)$$

$$z_{\rho_1} = 2.32$$

$$z_{\rho_2} = 2.98$$

Using Appendix 11, we convert the  $z_\rho$  values back into terms of the correlation coefficient and find, by interpolation,

$$\rho_1 \doteq 0.980$$

$$\rho_2 \doteq 0.999$$

We notice that the confidence limits are *not* symmetrical about  $r = 0.9907$  because of the lack of symmetry in the distribution of  $r$  about a value other than zero.

## 15.7 HYPOTHESES CONCERNING DIFFERENCES BETWEEN CORRELATION COEFFICIENTS

**Different Populations.** If correlation coefficients are calculated for two sets of the same pairs of variables which are separately sampled from two different populations, both assumed to be normal and bivariate for the variables in question, the following procedure may be used to test for significance of difference between the two correlation coefficients. Suppose that a group of 27 men (group a) and another group of 27 women (group b) are independently selected from a large number of job applicants. Both groups are given a manual dexterity test and a test of the applicant's ability to set rivets. The manual dexterity test is to be used to predict the applicant's ability to set rivets. The correlation coefficient for group a is  $r_a = 0.648$  and for group b,  $r_b = 0.500$ . Is the predictive value of the manual dexterity test for group a (males) greater than for group b (females)?

*Hypotheses:*

$$H_0: \rho_a - \rho_b = 0$$

$$H_1: \rho_a - \rho_b > 0$$

*Criterion of Significance:*

$$\alpha = 0.05$$

*Rejection Region:*

$$\text{From Appendix 3, } z_{0.05} = 1.645$$

*Standard Error:*

$$\sigma_{z_{r_a} - z_{r_b}} = \sqrt{\sigma_{z_{r_a}}^2 + \sigma_{z_{r_b}}^2} \quad (15-11)$$

and using Eq. (15-9), we have

$$\sigma_{z_{r_a} - z_{r_b}} = \sqrt{0.0411 + 0.0411} = 0.287$$

**Test Statistic:**

Using Appendix 11, we find  $z_{r_a} \doteq 0.77$  and  $z_{r_b} \doteq 0.55$ . Then

$$z = \frac{z_{r_a} - z_{r_b}}{\sigma_{z_{r_a} - z_{r_b}}} = \frac{0.77 - 0.55}{0.287} = 0.77$$

**Conclusion:**

Since  $z < z_{0.05}$ , we do not reject the null hypothesis, and we find that the predictive value of the test is not greater for men at the stated level of significance.

**Same Population.** A question that is commonly encountered is whether, for example,  $X_3$  is more highly correlated with  $X_1$  than is  $X_2$ , when all three variables have been drawn from the same population. Consider the following example. Two tests were given to a random sample of 27 male job applicants. The tests are being considered for use as predictors of applicants' ability to set rivets, variable 1.

Test number	Test name
2	Manual dexterity
3	Finger dexterity

The sample correlation coefficients are found to be

$$r_{12} = 0.648; \quad r_{13} = 0.597; \quad r_{23} = 0.642$$

Does test 2 do a better job as a predictor than test 3?

**Hypotheses:**

$$H_0: \rho_{12} - \rho_{13} = 0$$

$$H_1: \rho_{12} - \rho_{13} > 0$$

**Criterion of Significance:**

$$\alpha = 0.01$$

**Rejection Region:**

Using the  $t$  distribution with  $n - 3 = 24$  degrees of freedom, we find the rejection value of  $t$  to be 2.492.

**Test Statistic:**

Hotelling has suggested the statistic<sup>(7)</sup>

$$t = (r_{12} - r_{13}) \sqrt{\frac{(n-3)(1+r_{23})}{2(1-r_{12}^2-r_{13}^2-r_{23}^2+2r_{12}r_{13}r_{23})}}$$

<sup>(7)</sup> Harold Hotelling, "The Selection of Variates for Use in Prediction, with Some Comments on the General Problem of Nuisance Parameters," *Annals of Mathematical Statistics*, Vol. 11 (1940), p. 278.

Using our example data, we have

$$t = 0.408$$

*Conclusion:*

Since the calculated value of  $t$  fails to exceed the rejection value of  $t$ , we fail to reject the null hypothesis at the stated level of significance. Since we find no difference between the tests, it would probably be wise to use the less expensive of the two tests if a choice between them is to be made.

## 15.8 CORRELATION OF RANKED DATA

Sometimes statistical series are composed of items the exact magnitude of which cannot be ascertained, but which are ranked according to size. Thus the figures in column (2) of Table 15.2 show the ranks of 20 salesmen as determined by the sales manager and based upon their value to the concern. In column (3) their ranks are stated according to the results of two psychological tests. By comparing the ranks of the salesmen according to the

**TABLE 15.2: RANK OF 20 EMPLOYEES WITH RESPECT TO SELLING ABILITY AND RESULTS OF PSYCHOLOGICAL TESTS, AND COMPUTATION OF  $D^2$  FOR OBTAINING RANK CORRELATION COEFFICIENT**

<i>Salesmen</i>	<i>Rank by sales manager</i>	<i>Rank by two tests</i>	<i>D</i>	<i>D<sup>2</sup></i>
(1)	(2)	(3)	(4)	(5)
Millard	1	1	0	0
Prentice	2	6	-4	16
Borden	3	7	-4	16
McNulty	4	9	-5	25
Mattern	5	2	3	9
Peterson	6	10	-4	16
Rosoff	7	3	4	16
Haddad	8	5	3	9
Weingard	9	15	-6	36
Bochman	10	8	2	4
Gellers	11	4	7	49
Kelly	12	14	-2	4
Lyon	13	17	-4	16
Petty	14	18	-4	16
Kennedy	15	16	-1	1
Preston	16	12	4	16
Minnett	17	13	4	16
Tolan	18	11	7	49
Sullivan	19	19	0	0
Holman	20	20	0	0
<b>Total</b>	...	...	...	314

subjective criterion and according to the average score on the two tests, we see that the characteristics tested may have something to do with selling ability. For instance, Mr. Millard was first and Mr. Holman was last by either method of judging. Sometimes two or more items may be tied in rank; in this case each is given the average of the two ranks.<sup>(8)</sup> Thus if Borden and McNulty were tied for third and fourth, each would be given a rank of 3.5; if Borden, McNulty, and Mattern were tied for third, fourth, and fifth, each would be given a rank of 4.

A widely used rank correlation coefficient is Spearman's  $r_s$ .

$$r_s = 1 - \frac{6 \sum D^2}{n(n^2 - 1)} \quad (15-12)$$

where  $D$  is the difference between the pair of ranks. It can be shown (see Problem 5) that  $r_s$  gives the same result that could be obtained by applying the ordinary correlation coefficient to ranks. Referring to Table 15.2, we see that

$$r_s = 1 - \frac{6(314)}{(20)(400 - 1)} = +0.764$$

When  $n$  is large ( $n > 10$ ) the significance of the rank correlation coefficient may be tested in a manner analogous to Eq. (15-7) by forming the  $t$  statistic.<sup>(9)</sup>

$$t = r_s \sqrt{\frac{n-2}{1-r_s^2}}$$

using the  $t$  distribution with  $n - 2$  degrees of freedom (see Problem 5).

Spearman's  $r_s$  is known as a *distribution-free* (or nonparametric) measure of correlation, since strict assumptions are not made about the underlying distribution from which the sample observations were drawn. The price paid for relaxation of the assumptions inherent in ordinary correlation analysis is a loss of the power to reject a false null hypothesis. If a significant degree of correlation exists in the population, for a given level of alpha, it will require about 10 percent more observations using  $r_s$  to reject the null hypothesis than will be necessary using  $r$ . However,  $r_s$  is easier to compute than is  $r$ .

<sup>(8)</sup> If only a small proportion of the ranks are tied, this technique may be used in conjunction with Eq. (15-12). If a large proportion of the ranks are tied, it is advisable to apply a correction factor to Eq. (15-12). For a discussion of this matter as well as another rank correlation coefficient, Kendall's  $\tau$ , see: Sidney Siegel, *Nonparametric Statistics for the Behavioral Sciences*, (New York: McGraw-Hill Book Company, Inc., 1956), Chapter 9.

<sup>(9)</sup> Tables for testing Spearman's correlation coefficient when  $n \leq 10$  are available. For example, see Chemical Rubber Company, *Handbook of Tables for Probability and Statistics*, 1966, p. 330.



## PROBLEMS

1. Given the two sets of data below, show that although  $s_{1.2}^2$  and  $s_{3.2}^2$  are not the same,  $r_{12}$  and  $r_{32}$  are identical. Also show that  $r_{32} = r_{23}$ . Generalize your results.

$X_1$	$X_2$	$X_3$	$X_2$
50	1	0.5	1
200	2	2.0	2
250	3	2.5	3

2. Explain in words the meaning of  $r^2$  and  $1 - r^2$ . Why is expression (15-3) generally true? When will it not be true? (See Problem 7.)

3. Show that

$$t = \frac{b_{12}}{s_{b_{12}}} = r \sqrt{\frac{n-2}{1-r^2}}$$

Also show that this  $t$  statistic may be thought of as the geometric mean of the two  $t$  statistics associated with the  $b$ 's whose geometric mean defines  $r$ ; i.e.,

$$t = r \sqrt{\frac{n-2}{1-r^2}} = \sqrt{\frac{b_{12} b_{21}}{s_{b_{12}} s_{b_{21}}}}$$

Finally, show that the  $F$  ratio as defined by Eq. (15-8) is the square of the  $t$  statistic defined by Eq. (15-7).

4. Calculate one of the entries in Appendix 10.

5. If  $n$  values of  $X_1$  and  $X_2$  are in exact ranks, it can be shown that

$$\begin{aligned} \sum x_1^2 &= \sum x_2^2 = \frac{n(n^2 - 1)}{12} \\ \sum x_1 x_2 &= \frac{\sum x_1^2 + \sum x_2^2 - \sum D^2}{2} \end{aligned}$$

Show that Eqs. (15-4) and (15-12) are identical under these conditions. Also test

$$H_0: \rho_s = 0$$

$$H_1: \rho_s > 0$$

with  $\alpha = 0.05$  when  $r_s = +0.764$  and  $n = 20$ .

6. The following data were adopted from a study by Rolfe Wyer, "Learning Curve Helps Figure Profits," N.A.C.A. *Bulletin*, Vol. XXXV (December, 1953), pp. 402-502.

<i>Cumulative quantity</i> $X_2$	<i>Cumulative hours per unit</i> $X_1$
20	150
35	125
60	105
100	100
150	92
300	77
500	62
800	58
1500	47

- Plot the data.
- Plot the logarithms of the data.
- Fit a regression line to the original data and record  $r$ .
- Repeat part c, using the logarithms of the original data.
- Comment on the results of your calculations.

7. Show that  $r_{12}$  and  $r_{34}$  are meaningless. Explain.

$X_1$	$X_2$	$X_3$	$X_4$
3	1	2	1
5	2	2	2
		2	3

8. Show that the variables given in Problem 5, Chapter 11, are uncorrelated.

9. A functional relationship exists between the variables given below. Explain why  $r = 0$ .

$X_1$	$X_2$
4	-2
1	-1
0	0
1	1
4	2

## APPENDIX : Alternative Views of the Correlation Coefficient

The following are some alternative methods of stating the meaning of the simple linear correlation coefficient,  $r$ . In this appendix we will make use of the standard deviation

$$SD = \sqrt{\frac{\sum x^2}{n}}$$

**Covariance of Standardized Observations.** If we use  $SD_1$  and  $SD_2$  to standardize the variables  $X_1$  and  $X_2$  respectively

$$z_1 = \frac{X_1 - \bar{X}_1}{SD_1}; \quad z_2 = \frac{X_2 - \bar{X}_2}{SD_2}$$

then

$$r = \frac{\sum z_1 z_2}{n} \quad (\text{A15-1})$$

Equation (A15-1) is called the *product-moment* formula. It stresses the fact that  $r$  measures the extent to which two variables vary together when each has a mean of zero and a standard deviation of one.

**Slope of Standardized Observations.** In a similar manner, if we calculated the least-squares regression equation for the standardized observations and obtained the slope  $b_{z_1 z_2}$

$$r = b_{z_1 z_2} \quad (\text{A15-2})$$

Equation (A15-2) may be efficiently calculated by use of

$$r = b_{12} \frac{SD_2}{SD_1} \quad (\text{A15-3})$$

Therefore, in standard form, the regression equation may be written

$$z_{1(2)} = rz_2 \quad (\text{A15-4})$$

Using Eq. (A15-3), we can think of  $r$  as the slope  $b_{12}$ , adjusted for differences in variability in the two variables.

**Relative Reduction in Variance.** Since

$$r^2 = b_{12} \frac{\sum x_1 x_2}{\sum x_1^2} \quad (\text{A15-5})$$

and

$$\sum x_{1.2}^2 = \sum x_1^2 - \sum x_{1(2)}^2 = \sum x_1^2 - b_{12} \sum x_1 x_2$$

we may write Eq. (A15-5) as

$$r^2 = 1 - \frac{\sum x_{1.2}^2}{\sum x_1^2} \quad (\text{A15-6})$$

If we define  $(SD_{1.2})^2 = \sum x_{1.2}^2/n$  and  $(SD_1)^2 = \sum x_1^2/n$ , we may express Eq. (A15-6) as

$$r^2 = 1 - \frac{(SD_{1.2})^2}{(SD_1)^2} \quad (\text{A15-7})$$

Equation (A15-7) is a preferred expression by some statisticians because it states that  $r^2$  is the relative reduction in variance, when variance is measured about the regression line rather than about the mean of the dependent variable.

Sometimes a correlation coefficient is calculated that is known as the correlation coefficient *adjusted for degrees of freedom*. This coefficient is calculated by replacing the biased estimators of variance given in Eq. (A15-7) with the unbiased estimators  $s_{1.2}^2$  and  $s_1^2$ . The square of the coefficient is

$$\tilde{r}^2 = 1 - \frac{s_{1.2}^2}{s_1^2} = 1 - (1 - r^2) \left( \frac{n-1}{n-2} \right) \quad (\text{A15-8})$$

**Estimated Covariance Adjusted for Individual Estimated Variances.** Another method of expressing the correlation coefficient that stresses the covariability of the variables under the bivariate model is

$$r = \frac{s_{12}}{s_1 s_2} \quad (\text{A15-9})$$

where

$$s_{12} = \frac{\sum x_1 x_2}{n-1}$$

Also, it follows from Eq. (A15-9) that

$$r = \sqrt{\frac{s_{12}^2}{s_1^2 s_2^2}}$$

# 16

## Multiple and Partial Correlation and Regression

One method of reducing the magnitude of the error of the estimate and increasing the magnitude of the correlation coefficient is to use not one, but two or more independent variables in the estimating equation. Thus, in the example given in the previous two chapters, tensile strength might have been estimated on the basis of hardness and specific gravity rather than hardness alone.

### 16.1 A THREE-VARIABLE ILLUSTRATION

In this section we shall use an illustration from personnel testing that was alluded to in the last chapter.

A peg-board, or manual-dexterity, test and a copying, or finger-dexterity, test were administered to each of 27 aircraft trainees at San Diego, California. The statistical problem is how to weight these tests into a battery of tests which is in close agreement with a suitable criterion variable. In this case, the criterion variable is an objective test, the number of rivets set correctly per minute. Our variables then are

Dependent Variable:

$X_1$  Ability to set rivets

Independent Variables:

$X_2$  Manual-dexterity score

$X_3$  Finger-dexterity score

The values of all three variables are recorded in Table 16.1 and indicate achievement per unit of time.

The multiple estimating equation will take the form

$$\hat{X}_1 = a + b_{12.3}X_2 + b_{13.2}X_3 \quad (16-1)$$

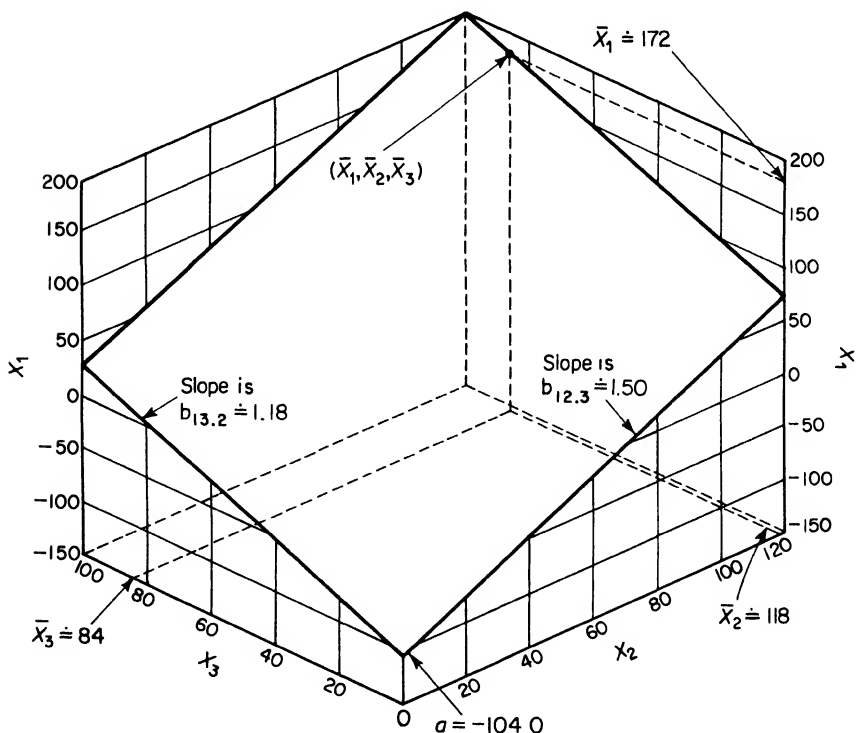
which, as we shall see, will be

$$\hat{X}_1 = -104.02 + 1.498X_2 + 1.182X_3 \quad (16-2)$$

The distinction between a simple and multiple estimating equation may be clarified if one thinks of the simple equation as describing a line and the multiple equation as describing a plane. Chart 16.1 shows part of the plane described by Eq. (16-2). The diagonal line on the right face of the solid indicates the change in  $X_1$  for each unit change in  $X_2$ , but a fixed value of  $X_3$ . Its slope is  $b_{12.3} = 1.498$ . This coefficient is called a *partial* regression coefficient, since the variable  $X_3$  is held fixed, or constant. The variable held constant is indicated by the number 3 after the dot in the subscript attached

**CHART 16.1: PLANE DESCRIBED BY MULTIPLE REGRESSION EQUATION:**

$$\hat{X}_1 = -104.0 + 1.50X_2 + 1.18X_3.$$



Source: Equation (16-2).

**TABLE 16.1: DATA FOR MULTIPLE CORRELATION OF MANUAL-DEXTERITY AND FINGER-DEXTERITY SCORE WITH CRITERION**

<i>Trainee number</i>	<i>Criterion <math>X_1</math></i>	<i>Manual dexterity <math>X_2</math></i>	<i>Finger dexterity <math>X_3</math></i>
1	230	135	107
2	81	93	67
3	100	108	81
4	212	138	93
5	216	123	81
6	156	116	86
7	201	119	86
8	194	112	96
9	164	128	80
10	166	116	86
11	146	125	78
12	196	114	89
13	202	128	84
14	203	129	80
15	201	125	99
16	195	120	86
17	180	126	92
18	174	136	95
19	120	104	82
20	198	116	76
21	189	112	80
22	184	109	85
23	174	113	75
24	168	113	87
25	143	104	69
26	131	103	65
27	130	125	84
Sum	4654	3190	2269
Mean	172.370	118.148	84.037
<i>Product of sums</i>	21,659,716 14,846,260 10,559,926	14,846,260 10,176,100 7,238,110	10,559,926 7,238,110 5,148,361
$\Sigma X_i$ $\bar{X}_i$	$\Sigma X_1$ $\bar{X}_1$	$\Sigma X_2$ $\bar{X}_2$	$\Sigma X_3$ $\bar{X}_3$
$\Sigma X_1 \Sigma X_2$ $\Sigma X_2 \Sigma X_1$ $\Sigma X_3 \Sigma X_1$	$(\Sigma X_1)^2$ $\Sigma X_2 \Sigma X_1$ $\Sigma X_3 \Sigma X_1$	$\Sigma X_1 \Sigma X_2$ $(\Sigma X_2)^2$ $\Sigma X_3 \Sigma X_2$	$\Sigma X_1 \Sigma X_3$ $\Sigma X_2 \Sigma X_3$ $(\Sigma X_3)^2$

Source: War Manpower Commission, courtesy of C. L. Shartle, Chief, Division of Occupational Analysis and Manning Tables.

The algebraic meaning of the calculated values is given below the double lines.

to the  $b_{12.3}$ . The diagonal line on the left face of the solid indicates the change in  $X_1$  for each unit change in  $X_3$ , but a fixed value of  $X_2$ . Its slope is  $b_{13.2} = 1.182$ . This coefficient, therefore, is also called a *partial* regression coefficient, since the variable  $X_2$  is held constant. Chart 16.1 shows as well that when  $X_2 = \bar{X}_2$  and  $X_3 = \bar{X}_3$ ,  $\hat{X}_1 = \bar{X}_1$ . Finally, the value of the intercept  $a$  is shown. This intercept is the computed value of  $\hat{X}_1$  when  $X_2$  and  $X_3$  are zero. One thing that Chart 16.1 does not show is the cloud of 27 observation points above and below the plane. This cloud is left to the reader's imagination. The plane has been fitted by the method of least squares, which, in a manner analogous to that described in Chapter 14, minimizes the sum of the squares of the vertical deviations of the observation points about the regression plane.

## 16.2 THE NORMAL EQUATIONS AND THEIR SOLUTION

Performing the mathematics necessary to minimize

$$\Sigma (X_1 - \hat{X}_1)^2$$

leads to the normal equations<sup>(1)</sup>

$$\left. \begin{array}{lcl} na + b_{12.3} \Sigma X_2 + b_{13.2} \Sigma X_3 = \Sigma X_1 & \text{I} \\ a \Sigma X_2 + b_{12.3} \Sigma X_2^2 + b_{13.2} \Sigma X_2 X_3 = \Sigma X_1 X_2 & \text{II} \\ a \Sigma X_3 + b_{12.3} \Sigma X_2 X_3 + b_{13.2} \Sigma X_3^2 = \Sigma X_1 X_3 & \text{III} \end{array} \right\} \quad (16-3)$$

By inserting the values given in Tables 16.1 and 16.2 in these equations and solving them simultaneously, using elementary algebraic techniques, we may obtain the values for the unknown coefficients. These equations, however, are more laborious to solve than is necessary. If we write the multiple estimating equation in deviation form similar to that which was used for the simple estimating equation, Eq. (14-25), the constant term will vanish, and

$$x_{1(23)} = b_{1.23}x_2 + b_{13.2}x_3 \quad (16-4)$$

<sup>(1)</sup> For three variables there are three normal equations; for four variables, four normal equations, and so on. A useful mnemonic device for remembering these equations is to view the  $i$ th normal equation as being the estimating equation multiplied by the coefficient of the  $i$ th unknown and then summed term by term. Thus, the coefficient of the first unknown,  $a$ , is 1, and we obtain

$$\Sigma [1(X_1 = a + b_{12.3}X_2 + b_{13.2}X_3)]$$

or

$$\Sigma X_1 = na + b_{12.3} \Sigma X_2 + b_{13.2} \Sigma X_3$$

In a similar manner

$$\Sigma [X_2(X_1 = a + b_{12.3}X_2 + b_{13.2}X_3)]$$

or

$$\Sigma X_1 X_2 = a \Sigma X_2 + b_{12.3} \Sigma X_2^2 + b_{13.2} \Sigma X_2 X_3$$

and so on for the remaining equation.



TABLE 16.2: COMPUTATION OF VALUES FOR MULTIPLE CORRELATION OF MANUAL-DEXTERITY AND FINGER-DEXTERITY SCORES WITH CRITERION

Trainee	$X_1^2$	$X_1X_2$	$X_1X_3$	$X_2^2$	$X_2X_3$	$X_3^2$
1	52,900	31,050	24,610	18,225	14,455	11,449
2	6,561	7,533	5,427	8,649	6,231	4,489
.	.	.	.	.	.	.
.	.	.	.	.	.	.
26	17,161	13,493	8,515	10,609	6,695	4,225
27	16,900	16,250	10,920	15,625	10,500	7,056
Sum of squares or products	836,924	556,637	396,486	380,040	269,820	193,021
Correction term	802,211.70	549,861.48	391,108.38	376,892.59	268,078.15	190,680.04
Variation or covariation	34,712.30	6,775.52	5,377.62	3,147.41	1,741.85	2,340.96
$\Sigma X_i X_j$	$\Sigma X_1^2$	$\Sigma X_1 X_2$	$\Sigma X_1 X_3$	$\Sigma X_2^2$	$\Sigma X_2 X_3$	$\Sigma X_3^2$
$\frac{\Sigma X_i \Sigma X_j}{n}$	$\frac{(\Sigma X_1)^2}{n}$	$\frac{\Sigma X_1 \Sigma X_2}{n}$	$\frac{\Sigma X_1 \Sigma X_3}{n}$	$\frac{(\Sigma X_2)^2}{n}$	$\frac{\Sigma X_2 \Sigma X_3}{n}$	$\frac{(\Sigma X_3)^2}{n}$
$\Sigma x_i x_j$	$\Sigma x_1^2$	$\Sigma x_1 x_2$	$\Sigma x_1 x_3$	$\Sigma x_2^2$	$\Sigma x_2 x_3$	$\Sigma x_3^2$

Source: Table 16.1. The algebraic meaning of the calculated values is given below the double lines.

which leads to two normal equations:

$$\left. \begin{aligned} b_{1.23} \sum x_2^2 + b_{13.2} \sum x_2 x_3 &= \sum x_1 x_2 & \text{II}' \\ b_{12.3} \sum x_2 x_3 + b_{13.2} \sum x_3^2 &= \sum x_1 x_3 & \text{III}' \end{aligned} \right\} \quad (16-5)$$

Then, if we divide both sides of I by  $n$ , we obtain

$$a = \bar{X}_1 - b_{12.3} \bar{X}_2 - b_{13.2} \bar{X}_3 \quad (16-6)$$

The  $b$  coefficients have the same meaning and numerical value whether the equations are in original or deviation form. They are, however, simpler to obtain by using the fewer number of equations resulting from expressing the estimating equation in deviation form.

The measures of variation and covariation are found in the same manner as that given in Chapter 14 and are displayed in Table 16.2. Once again, they are found by computing the matrix of sums of squares and cross products and subtracting from it a matrix of correction terms to give a matrix of variation and covariation. Symbolically, and again omitting the redundant terms below the main diagonal of the matrix, we have

$$\begin{array}{c} \begin{matrix} (1) & (2) & (3) \\ (1) \left[ \begin{matrix} \sum X_1^2 & \sum X_1 X_2 & \sum X_1 X_3 \\ (2) \left[ \begin{matrix} & \sum X_2^2 & \sum X_2 X_3 \\ (3) \left[ \begin{matrix} & & \sum X_3^2 \end{matrix} \end{matrix} \right] \end{matrix} \right] \end{matrix} \end{array} - \begin{array}{c} \begin{matrix} (1) & (2) & (3) \\ (1) \left[ \begin{matrix} \frac{(\sum X_1)^2}{n} & \frac{\sum X_1 \sum X_2}{n} & \frac{\sum X_1 \sum X_3}{n} \\ (2) \left[ \begin{matrix} & \frac{(\sum X_2)^2}{n} & \frac{\sum X_2 \sum X_3}{n} \\ (3) \left[ \begin{matrix} & & \frac{(\sum X_3)^2}{n} \end{matrix} \end{matrix} \right] \end{matrix} \right] \end{matrix} \end{array} \end{array} \end{array}$$

$$- \begin{array}{c} \begin{matrix} (1) & (2) & (3) \\ (1) \left[ \begin{matrix} \sum x_1^2 & \sum x_1 x_2 & \sum x_1 x_3 \\ (2) \left[ \begin{matrix} & \sum x_2^2 & \sum x_2 x_3 \\ (3) \left[ \begin{matrix} & & \sum x_3^2 \end{matrix} \end{matrix} \right] \end{matrix} \right] \end{matrix} \end{array}$$

Using the data given in Table 16.2, we find the normal equations in deviation form to be

$$\begin{aligned} 3147.41b_{12.3} + 1741.85b_{13.2} &= 6775.52 & \text{II}' \\ 1741.85b_{12.3} + 2340.96b_{13.2} &= 5377.62 & \text{III}' \end{aligned}$$

Now it is quite easy to perform the algebra necessary to find the two partial regression coefficients,<sup>(2)</sup> and our estimating equation in deviation form is

<sup>(2)</sup> Multiply II' by  $\sum x_2 x_3 / \sum x_3^2 = 1741.85/3147.41 = 0.5534233$  and subtract the resulting equation from III' to obtain  $b_{13.2}$ .

$$\begin{array}{rcl} 1741.85b_{12.3} + 2340.96b_{13.2} & = & 5377.62 & \text{III}' \\ 1741.85b_{12.3} + 963.98b_{13.2} & = & 3749.73 & \text{II}' \times (0.5534233) \\ \hline 1376.98b_{13.2} & = & 1627.89 \\ b_{13.2} & = & 1.18222 \end{array}$$

(Continued on p. 234)

$$x_{1(23)} = 1.49846x_2 + 1.18222x_3$$

To convert this equation into original form, we must find  $a$ , the intercept. Using Eq. (16-6), we have

$$\begin{aligned} a_{1(23)} &= 172.370 - 1.49846(118.148) - 1.18222(84.037) \\ &= -104.02 \end{aligned}$$

The estimating equation may now be written after rounding as is given by Eq. (16-2).

### 16.3 SOURCES OF VARIATION

In a manner strictly analogous to simple regression, we now proceed to partition total variation  $\sum x_1^2$  into explained  $\sum x_{1(23)}^2$  and unexplained  $\sum x_{1.23}^2$  components. The basic extension involves the fact that we now regard *both* the variables  $X_2$  and  $X_3$  in the computation of explained variation. Table 16.3 summarizes these calculations.

TABLE 16.3: TOTAL, EXPLAINED, AND UNEXPLAINED VARIATION

Source of variation	Symbol	Computational formula	Amount of variation (example values)
Total	$\sum x_1^2$	$\sum x_1^2 = \sum X_1^2 - \frac{(\sum X_1)^2}{n}$	34,712
Explained	$\sum x_{1(23)}^2$	$\sum x_{1(23)}^2 = b_{12.3} \sum x_1 x_2 + b_{13.2} \sum x_1 x_3$	16,510
Unexplained	$\sum x_{1.23}^2$	$\sum x_{1.23}^2 = \sum x_1^2 - \sum x_{1(23)}^2$	18,202

The standard error of estimate is given by

$$\begin{aligned} s_{1.23} &= \sqrt{\frac{\sum x_{1.23}^2}{n - m}} \\ &= \sqrt{\frac{18,202}{24}} = \sqrt{758.4} = 27.54 \end{aligned} \quad (16-7)$$

where  $m$  is the number of constants in the estimating equation. In this example, there are  $n - 3$  degrees of freedom because the estimating equation has 3 constants. For a four-variable problem there would be  $n - 4$  degrees of freedom, and so on.

Then substitute  $b_{13.2} = 1.18222$  into either II' or III' to obtain  $b_{12.3} = 1.49846$ . It is advisable to check the calculations by insertion of the calculated values of the slopes into both of the normal equations.

## 16.4 MULTIPLE CORRELATION

The multiple correlation coefficient is calculated in a manner strictly analogous to that given for the simple correlation coefficient. Thus,

$$r_{1(23)} = \sqrt{\frac{\text{Explained variation}}{\text{Total variation}}} = \sqrt{\frac{\sum x_{1(23)}^2}{\sum x_1^2}} \quad (16-8)$$

and for this problem

$$r_{1(23)} = \sqrt{\frac{16,510}{34,712}} = \sqrt{0.4756} = 0.6897$$

Notice that no sign is affixed to the coefficient of multiple correlation, since the different  $b$  coefficients do not necessarily have the same sign. The meaning of the multiple correlation coefficient will be elaborated upon in the next section.

## 16.5 PARTIAL CORRELATION AND RELATIONSHIPS BETWEEN CORRELATION COEFFICIENTS

Associated with any multiple regression and correlation problem is a set of simple correlation problems which describes the simple relationships between pairs of variables. For example, if we were so disposed, we could calculate for every pair of variables simple regression equations as well as simple correlation coefficients.

The partial regression coefficient  $b_{12.3}$  may be thought of as the slope of  $x_1$  on  $x_{2.3}$ . In this sense the effect of variable 3 is "held constant." In an analogous manner there is the partial correlation coefficient  $r_{12.3}$ , which is the correlation between  $x_1$  and  $x_2$  with variable 3 held constant. We now explore the meaning of these partial correlation coefficients as well as other correlation coefficients associated with a multiple correlation problem.

Partial correlation coefficients show the net correlation between each independent variable and the dependent variable. In this sense they show the relative importance of each independent variable. It is enlightening to compare the concepts of partial correlation with those of simple and multiple correlation.

The simple correlation coefficient  $r_{12}$  is the square root of the proportion of total variation that has been explained by use of the equation  $\hat{X}_1 = a + b_{12}X_2$  or  $x_{1(2)} = b_{12}x_2$ . Also, it may be regarded as the simple correlation between  $X_1$  and  $\hat{X}_1$  or between  $x_1$  and  $x_{1(2)}$ .

In our example the multiple correlation coefficient  $r_{1(23)}$  is the square root of the proportion of variation that has been explained by use of the equation  $\hat{X}_1 = a + b_{12.3}X_2 + b_{13.2}X_3$  or  $x_{1(23)} = b_{12.3}x_2 + b_{13.2}x_3$ . Also, it may be

regarded as the simple correlation between  $X_1$  and  $\hat{X}_1$  or between  $x_1$  and  $x_{1(23)}$  as defined directly above.

The partial correlation coefficient  $r_{12.3}$  may be regarded as the simple correlation between the variables 1 and 2 when each of these has been adjusted for the effect of variable 3; that is, the simple correlation between  $x_{1.3}$  and  $x_{2.3}$ . (Likewise,  $r_{13.2}$  is the simple correlation between  $x_{1.2}$  and  $x_{3.2}$ .) For computational purposes it is perhaps the simplest to regard  $r_{12.3}^2$  as the *net* proportion of variation explained<sup>(3)</sup> by variable 2; that is, the proportion of variation unexplained after use of variable 3, which was explained by variable 2. Thus,  $r_{12.3}$  is obtained by finding the *increase* in the explained variation brought about by use of the equation  $x_{1(23)} = b_{12.3}x_2 + b_{13.2}x_3$  instead of the equation  $x_{1(3)} = b_{13}x_3$ , dividing this quantity by the amount of variation still to be explained after using the latter equation alone, and extracting the square root. The coefficient  $r_{13.2}$  is obtained in an analogous manner.

$$\left. \begin{aligned} r_{12.3}^2 &= \frac{\sum x_{12.3}^2}{\sum x_{1.3}^2} = \frac{\sum x_{1(23)}^2 - \sum x_{1(3)}^2}{\sum x_1^2 - \sum x_{1(3)}^2} \\ r_{13.2}^2 &= \frac{\sum x_{13.2}^2}{\sum x_{1.2}^2} = \frac{\sum x_{1(23)}^2 - \sum x_{1(2)}^2}{\sum x_1^2 - \sum x_{1(2)}^2} \end{aligned} \right\} \quad (16-9)$$

Note that for both the multiple coefficient and the partial coefficients the subscript to  $r$  is always the same as the subscript to  $\sum x^2$  in the numerator. For partial coefficients, the variable that has been held constant follows the decimal point in the subscript to the coefficient and also in the formula itself. Thus,

$$r_{12}^2 = \frac{\sum x_{1(2)}^2}{\sum x_1^2}$$

and we affix .3 everywhere to obtain

$$r_{12.3}^2 = \frac{\sum x_{12.3}^2}{\sum x_{1.3}^2}$$

Note also that the full formula is the same as the multiple correlation formula, except that both the numerator and the denominator are adjusted for variation explained by the variable held constant. Thus

$$r_{1(23)}^2 = \frac{\sum x_{1(23)}^2}{\sum x_1^2}$$

but

$$r_{12.3}^2 = \frac{\sum x_{1(23)}^2 - \sum x_{1(3)}^2}{\sum x_1^2 - \sum x_{1(3)}^2}$$

The signs of  $r_{12.3}$  and  $r_{13.2}$ , respectively, are the same as those of  $b_{12.3}$  and  $b_{13.2}$ .

<sup>(3)</sup> Another way of looking at the net proportion of variation explained is instructive. Just as

$$\sum x_{1(3)}^2 = b_{13} \sum x_1 x_3 \quad \text{and} \quad \sum x_{1(3)}^2 = b_{13} \sum x_1 x_3$$

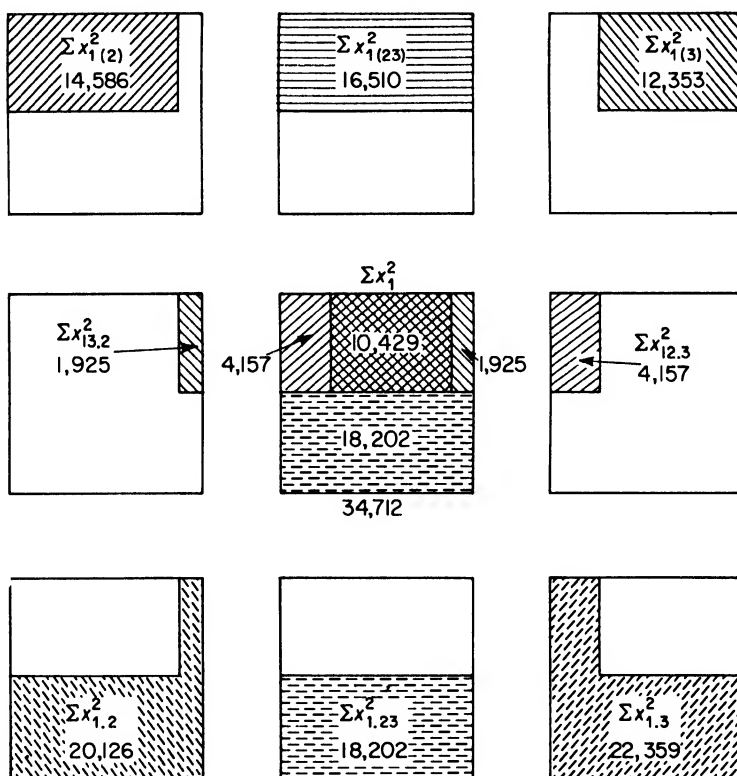
so  $\sum x_{12.3}^2 = b_{12.3} \sum x_{1.3} x_{2.3}$  and  $\sum x_{13.2}^2 = b_{13.2} \sum x_{1.2} x_{3.2}$

This is the main approach used in the Doolittle method (see the appendix to this chapter).

A summary of the formulas for simple, multiple, and partial correlation at this point may be useful and is given in Table 16.4.

The relationships among the coefficients of simple, multiple, and partial determination will perhaps be still further clarified by reference to Chart 16.2, in which the total variation is partitioned into the various types of explained and unexplained variation. One important principle is brought out by the chart. There is a great deal of overlapping between  $\sum x_{1(2)}^2$  and  $\sum x_{1(3)}^2$  (10,429 in this case). This is the amount of variation explained jointly by variables 2 and 3. As the chart indicates,  $\sum x_{1(23)}^2 = \sum x_{12.3}^2 + \sum x_{13.2}^2 +$

**CHART 16.2: DIAGRAMMATIC REPRESENTATION OF SOURCES OF VARIATION IN DEPENDENT VARIABLE WITH TWO INDEPENDENT VARIABLES.**



Source: Table 16.1. Notice that, in terms of set notation

$$\sum x_{1(2)}^2 \cup \sum x_{1(3)}^2 = \sum x_{1(23)}^2$$

and if

$$\sum x_1^2 > \sum x_{12.3}^2 + \sum x_{13.2}^2$$

then  $\sum x_{1(23)}^2 = \sum x_{12.3}^2 + \sum x_{13.2}^2 + (\sum x_{1(2)}^2 \cap \sum x_{1(3)}^2)$

where  $\sum x_{1(2)}^2 \cap \sum x_{1(3)}^2$  is the joint-explained variation.

TABLE 16.4: SUMMARY OF THE FORMULAS FOR COEFFICIENTS OF SIMPLE, MULTIPLE, AND PARTIAL DETERMINATION

Type of relationship	Symbol	COEFFICIENT OF DETERMINATION	
		Formula	
Simple	$r_{12}^2$	$\frac{\sum x_{1(2)}^2}{\sum x_1^2} = \frac{b_{12} \sum x_1 x_2}{\sum x_1^2}$	$\frac{2.15272(6775.52)}{34,712.30} = \frac{14,585.80}{34,712.30} = 0.4202$
Simple	$r_{13}^2$	$\frac{\sum x_{1(3)}^2}{\sum x_1^2} = \frac{b_{13} \sum x_1 x_3}{\sum x_1^2}$	$\frac{2.29719(5377.63)}{34,712.30} = \frac{12,353.44}{34,712.30} = 0.3559$
Simple	$r_{23}^2$	$\frac{\sum x_{2(3)}^2}{\sum x_2^2} = \frac{b_{23} \sum x_2 x_3}{\sum x_2^2}$	$\frac{0.7441(1741.85)}{3147.41} = \frac{1296.1}{3147.41} = 0.4118$
Multiple	$r_{1(23)}^2$	$\frac{\sum x_{1(23)}^2}{\sum x_1^2} = \frac{b_{12.3} \sum x_1 x_2 + b_{13.2} \sum x_1 x_3}{\sum x_1^2}$	$\frac{1.49846(6775.52) + 1.18222(5377.63)}{34,712.30} = 0.4756$
Partial	$r_{12.3}^2$	$\frac{\sum x_{12.3}^2}{\sum x_{1.3}^2} = \frac{\sum x_{1(23)}^2 - \frac{\sum x_{1(2)}^2 \sum x_{1(3)}^2}{\sum x_1^2}}{\sum x_1^2 - \frac{\sum x_{1(2)}^2 \sum x_{1(3)}^2}{\sum x_1^2}}$	$\frac{16,510.4 - 12,353.44}{34,712.30 - 12,353.44} = \frac{4156.96}{22,358.86} = 0.1859$
Partial	$r_{13.2}^2$	$\frac{\sum x_{13.2}^2}{\sum x_{1.2}^2} = \frac{\sum x_{1(23)}^2 - \frac{\sum x_{1(2)}^2 \sum x_{1(3)}^2}{\sum x_1^2}}{\sum x_1^2 - \frac{\sum x_{1(2)}^2 \sum x_{1(3)}^2}{\sum x_1^2}}$	$\frac{16,510.4 - 14,585.80}{34,712.30 - 14,585.80} = \frac{1924.60}{20,126.50} = 0.0965$

joint explained variation. The total variation explained by variables 2 and 3 together is equal to the net amount explained by variable 2 plus the net amount explained by variable 3 plus the amount explained jointly by them. It is this duplication that prevents the coefficient of multiple determination from being the sum of the two coefficients of simple determination. The chart also shows that

$$\sum x_{1(23)}^2 = \sum x_{1(2)}^2 + \sum x_{13.2}^2 = \sum x_{1(3)}^2 + \sum x_{12.3}^2$$

i.e.,<sup>(4)</sup> the total variation explained by variables 2 and 3 together is equal to the gross amount explained by variable 2 plus the net amount explained by variable 3, or the gross amount explained by variable 3 plus the net amount explained by variable 2. Still another way of looking at Chart 16.2 is that  $\sum x_{1(23)}^2 = \sum x_{1(2)}^2 + \sum x_{1(3)}^2 - \text{joint explained variation}$ ; i.e., the total variation explained by variables 2 and 3 together is equal to the gross amount explained by variable 2 plus the gross amount explained by variable 3 minus the amount explained jointly by them. For given values of  $r_{12}$  and  $r_{13}$  having the same sign, the duplication between  $\sum x_{1(2)}^2$  and  $\sum x_{1(3)}^2$  will be large if  $r_{23}$  is large and positive and will become smaller as  $r_{23}$  gets closer to zero. It follows that for given values of  $r_{12}$  and  $r_{13}$  having the same sign,<sup>(5)</sup> the smaller (algebraically) the value of  $r_{23}$ , the larger will be the value<sup>(6)</sup> of  $r_{1(23)}$ . Consequently, if a number of independent variables are available for use in multiple correlation with the dependent variable, and if each of these is correlated positively with the dependent variable, one should select those variables that have large correlation with the dependent variable and small correlation (algebraically) with each other.<sup>(7)</sup>

<sup>(4)</sup> Accordingly,

$$\sum x_{1(23)}^2 = b_{12} \sum x_1 x_2 + b_{13.2} \sum x_{1.2} x_{2.3}$$

or

$$\sum x_{1(23)}^2 = b_{13} \sum x_1 x_3 + b_{12.3} \sum x_{1.3} x_{2.3}$$

This method of computation is the main approach used in the abbreviated Doolittle method.

<sup>(5)</sup> +0.2 and -0.2 are both smaller *numerically* than is -0.3, but they are both larger *algebraically* than -0.3; +0.4 is larger than any of these numbers, both numerically and algebraically.

<sup>(6)</sup> This fact is apparent from a consideration of the relationship

$$r_{1(23)}^2 = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}$$

In applying this formula it must be remembered that for a given value of  $r_{12}$  and  $r_{13}$  only a certain range of values for  $r_{23}$  is possible when one is dealing with real numbers.

<sup>(7)</sup> If  $r_{12}$  and  $r_{13}$  are of opposite signs,  $\sum x_{1(23)}^2 > \sum x_{1(2)}^2 + \sum x_{1(3)}^2$ , and the larger, algebraically, the value of  $r_{23}$ , the larger will be the value of  $r_{1(23)}$ . If the sign of the correlation with the independent variable is negative for any of the variables, one should change (one at a time) the sign of the coefficient of correlation of each such variable with the dependent variable and each of the other independent variables. This process, called reflection, results in each independent variable's showing a positive correlation with the dependent variable. After reflection, the relationships among the different measures of variation and the different correlation coefficients are as described above.



## 16.6 ALTERNATIVE APPROACHES TO REGRESSION AND CORRELATION

Instead of solving simultaneous equations, one can first obtain  $b_{12}$ ,  $b_{13}$ ,  $b_{23}$  and other *zero-order* regression coefficients. From these,  $b_{13.2}$ ,  $b_{12.3}$ , and other *first-order* coefficients can quickly be obtained. The procedure can be extended indefinitely to obtain coefficients of any order,<sup>(8)</sup> though it is not so fast as the abbreviated Doolittle method if there are more than four variables.

Let 1 denote the dependent variable, 2 denote the "active" independent variable, 3, 4, . . . denote the "passive" independent variables (those held constant). Then

$$\left. \begin{aligned} b_{12.3} &= \frac{b_{12} - b_{13}b_{32}}{1 - b_{23}b_{32}} \\ b_{12.34} &= \frac{b_{12.4} - b_{13.4}b_{32.4}}{1 - b_{23.4}b_{32.4}} \end{aligned} \right\} \quad (16-10)$$

and so on.

Partial correlation coefficients can be obtained from partial regression coefficients:

$$\left. \begin{aligned} r_{12.3} &= \sqrt{b_{12.3}b_{21.3}} \\ r_{12.34} &= \sqrt{b_{12.34}b_{21.34}} \end{aligned} \right\} \quad (16-11)$$

and so on.

Also, partial correlation coefficients can be obtained from correlation coefficients of lower order:

$$\left. \begin{aligned} r_{12.3} &= \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2} \sqrt{1 - r_{23}^2}} \\ r_{12.34} &= \frac{r_{12.3} - r_{13.4}r_{23.4}}{\sqrt{1 - r_{13.4}^2} \sqrt{1 - r_{23.4}^2}} \end{aligned} \right\} \quad (16-12)$$

and so on.

Finally, the partial regression coefficients are sometimes adjusted for differences in the units measurement of the dependent variable and the active independent variable. The adjusted coefficients are often called *beta weights*, or *standard* partial regression coefficients. These beta weights, like correlation coefficients, are free from the influence of units of measurement and, therefore,

<sup>(8)</sup> This topic is fully explained and illustrated in Dudley J. Cowden, "A Procedure for Computing Regression Coefficients," *Journal of the American Statistical Association*, Vol. 53 (March, 1958), pp. 144-150.

are comparable for different sets of variables. For example,

$$\left. \begin{aligned} \beta_{12} &= b_{12} \frac{s_2}{s_1} \\ \beta_{12.3} &= b_{12.3} \frac{s_2}{s_1} \\ \beta_{12.34} &= b_{12.34} \frac{s_2}{s_1} \end{aligned} \right\} \quad (16-13)$$

and so on. In each case the beta weights are the ordinary regression coefficients that could be calculated if all of the variables in the regression problem had been standardized prior to the calculation of the estimating equation. Thus,  $z_{1(23)} = \beta_{12.3}z_2 + \beta_{13.2}z_3$ .

## 16.7 TESTING SIGNIFICANCE

Analysis of variance can be used for testing significance not only of simple, but also of multiple and partial correlation, as well as the constants in the estimating equation. Reference to Chart 16.2 will help one in understanding Table 16.5. Further explanation concerning degrees of freedom is helpful. There are  $n = 27$  sets of observations and  $m = 3$  constants; one degree of freedom is lost through use of the mean, so that total variation has  $n - 1 = 26$  degrees of freedom. Two additional ( $m - 1$ ) degrees of freedom are lost through use of the constants  $b_{12.3}$  and  $b_{13.2}$ . (It would be duplication to count both  $\bar{X}_1$  and  $a$ .) Note that degrees of freedom are additive in the same way as are the amounts of variation; thus for  $\sum x_{1.23}^2$

$$(n - 1) - (m - 1) - 26 - 2 = 24$$

or

$$n - m = 27 - 3 = 24$$

**Multiple Correlation.**<sup>(9)</sup> Let us test the null hypothesis that  $\rho_{1(23)} = 0$  against the alternative that  $\rho_{1(23)} \neq 0$ , at  $\alpha = 0.05$ . Then the  $F$  ratio is

$$F_{1(23)} = \frac{s_{1(23)}^2}{s_{1.23}^2} = \frac{\sum x_{1(23)}^2 / (m - 1)}{\sum x_{1.23}^2 / (n - m)} \quad (16-14)$$

and can be written directly in terms of the multiple correlation coefficient.

$$F_{1(23)} = \frac{r_{1(23)}^2 / (m - 1)}{(1 - r_{1(23)}^2) / (n - m)} \quad (16-15)$$

<sup>(9)</sup> This is also a test of the hypothesis  $B_{12.3} = B_{13.2} = 0$ . In the case of  $m$  variables it is a test of the hypothesis  $B_{12.34 \dots m} = B_{13.24 \dots m} = \dots = B_{1m.23 \dots m-1} = 0$  against the alternative that *not all* of the coefficients are zero. Thus, it is a test of the general significance of regression.

Using the data of Table 16.5 and Eq. (16-14), we find that

$$F_{1(23)} = \frac{16,510/2}{18,202/24} = \frac{8255}{758} = 10.89$$

The student may verify that the same value of  $F_{1(23)}$  can be obtained from Eq. (16-15). From Appendix 5 we find that with  $\alpha = 0.05$  and  $\nu_1 = 2$ ,  $\nu_2 = 24$ , the upper rejection value for  $F$  is 3.40. Notice that we are using only the upper rejection value of  $F$ , since the null hypothesis will be rejected when  $F$  is "large." Since the calculated value of  $F_{1(23)}$  exceeds the rejection value, we reject the null hypothesis and conclude that the multiple correlation coefficient is significant.<sup>(10)</sup>

**Partial Correlation.** Let us test

$$H_0: \rho_{12.3} = 0$$

$$H_1: \rho_{12.3} > 0$$

**TABLE 16.5: ANALYSIS OF VARIANCE FOR TESTING SIGNIFICANCE OF RELATIONSHIP BETWEEN MANUAL-DEXTERITY SCORE AND FINGER-DEXTERITY SCORE, AND CRITERION**

Source of variation	Symbol	Amount of variation	Degrees of freedom	Estimated variance*
Total	$\Sigma x_1^2$	34,712	$n - 1 = 26$	...
Less:				
Explained by $X_2$ and $X_3$	$\Sigma x_{1(23)}^2$	16,510	$m - 1 = 2$	$s_{1(23)}^2 = 8255$
Unexplained	$\Sigma x_{1.23}^2$	18,202	$n - m = 24$	$s_{1.23}^2 = 758$
Explained by $X_2$ and $X_3$	$\Sigma x_{1(23)}^2$	16,510	2	...
Less:				
Gross explained by $X_2$	$\Sigma x_{1(2)}^2$	12,353	1	...
Net explained by $X_2$	$\Sigma x_{12.3}^2$	4,157	1	$s_{12.3}^2 = 4157$
Explained by $X_2$ and $X_3$	$\Sigma x_{1(23)}^2$	16,510	2	...
Less:				
Gross explained by $X_3$	$\Sigma x_{1(3)}^2$	14,585	1	...
Net explained by $X_3$	$\Sigma x_{13.2}^2$	1,925	1	$s_{13.2}^2 = 1925$

\* Each of these is an unbiased estimate of the population variance if the population is uncorrelated.

<sup>(10)</sup> Confidence limits for the multiple correlation coefficient may be set. The technique, however, is beyond the scope of this text. See: K. H. Kramer, "Tables for Constructing Confidence Limits on the Multiple Correlation Coefficient," *Journal of the American Statistical Association*, Vol. 58 (December, 1963), pp. 1082-1085.

with  $\alpha = 0.025$ . Then the  $F$  ratio is

$$F_{12.3} = \frac{s_{12.3}^2}{s_{1.23}^2} = \frac{\sum x_{12.3}^2 / (m - k - 1)}{\sum x_{1.23}^2 / (n - m)} \quad (16-16)$$

where  $k$  is the number of "passive" variables, i.e., the number of variables held constant. Equation (16-16) can be written directly in terms of the partial correlation coefficient:

$$F_{12.3} = \frac{r_{12.3}^2 / (m - k - 1)}{(1 - r_{12.3}^2) / (n - m)} \quad (16-17)$$

Using Eq. (16-16) and the data of Table 16.5, we find that

$$F_{12.3} = \frac{4157/1}{18,202/24} = 5.484$$

From Appendix 5 we find that with  $2\alpha = 0.05$  and  $\nu_1 = 1$ ,  $\nu_2 = 24$ , the rejection value for  $F$  is 4.26. Since  $F_{12.3} > 4.26$ , we reject the null hypothesis.

To test the significance of  $r_{13.2}$  we form

$$F_{13.2} = \frac{s_{13.2}^2}{s_{1.23}^2} = \frac{\sum x_{13.2}^2 / (m - k - 1)}{\sum x_{1.23}^2 / (n - m)} \quad (16-18)$$

or

$$F_{13.2} = \frac{r_{13.2}^2 / (m - k - 1)}{(1 - r_{13.2}^2) / (n - m)} \quad (16-19)$$

Using Eq. (16-18) and the data of Table 16.5, we find

$$F_{13.2} = \frac{1925/1}{18,202/24} = 2.540$$

which is not greater than the upper rejection value of  $F$  at  $2\alpha = 0.05$ ,  $\nu_1 = 1$  and  $\nu_2 = 24$  (i.e.,  $2.540 < 4.26$ ). We conclude that  $r_{13.2}$  is not significant.

Just as a significance test for  $r_{12}$  was shown to be equivalent to a significance test for  $b_{12}$  in Sec. 15.5, so it can be shown that testing the significance of  $r_{12.3}$  is equivalent to testing the significance of  $b_{12.3}$  and, similarly, testing the significance of  $r_{13.2}$  tests the significance of  $b_{13.2}$ . Thus,  $b_{12.3}$  has been found to be greater than zero, whereas  $b_{13.2}$  has not, according to our previously discussed tests for  $r_{12.3}$  and  $r_{13.2}$ .<sup>(11)</sup>

The reason  $\sum x_{1.23}^2 = 18,202$  is used for the unexplained variation for both  $F_{12.3}$  and  $F_{13.2}$  may not be obvious. Table 16.5, in which the total variation

<sup>(11)</sup> Tests of other hypotheses concerning  $b_{13.1}$  and  $b_{12.1}$  can be conducted by comparing the difference between the calculated value of the partial regression coefficient and a hypothetical value with the estimated standard error of the partial regression coefficient. The method follows that illustrated for simple regression coefficients in Sec. 14.5. A  $t$  statistic results with  $n - m$  degrees of freedom. Tests of hypotheses concerning the intercept are carried out in a similar manner with  $n - m$  degrees of freedom. The standard errors of the intercept and partial regression coefficients are defined in an appendix to this chapter.

is analyzed into various additive components, may make this clear. The following equalities are easily seen by inspection of Chart 16.2.

$$\Sigma x_1^2 = \Sigma x_{1(2)}^2 + \Sigma x_{13.2}^2 + \Sigma x_{1.23}^2 = 14,585 + 1925 + 18,202 = 34,712$$

$$\Sigma x_1^2 = \Sigma x_{1(3)}^2 + \Sigma x_{12.3}^2 + \Sigma x_{1.23}^2 = 12,353 + 4157 + 18,202 = 34,712$$

In nonmathematical language we may say that total variation is made up of the gross amount explained by one independent variable plus the net amount explained by the other independent variable plus the amount not explained by the two taken together. Notice also

$$\Sigma x_{1.23}^2 = \Sigma x_1^2 - \Sigma x_{1(2)}^2 - \Sigma x_{13.2}^2 = \Sigma x_{1.2}^2 - \Sigma x_{13.2}^2$$

$$\Sigma x_{1.23}^2 = \Sigma x_1^2 - \Sigma x_{1(3)}^2 - \Sigma x_{12.3}^2 = \Sigma x_{1.3}^2 - \Sigma x_{12.3}^2$$

In other words, if we subtract the amount of variation we have explained from the amount we are trying to explain, we have the amount still unexplained  $\Sigma x_{1.23}^2$ .

The significance of partial correlation coefficients (and hence the significance of the partial regression coefficients) can also be tested by using a  $t$  statistic in a manner analogous to that described in Sec. 15.5 for the simple correlation coefficient:

$$\left. \begin{aligned} t_{12.3} &= r_{12.3} \sqrt{\frac{n-k-2}{1-r_{12.3}^2}} \\ t_{13.2} &= r_{13.2} \sqrt{\frac{n-k-2}{1-r_{13.2}^2}} \end{aligned} \right\} \quad (16-20)$$

and the test may be conducted directly by using Appendix 10 with  $\nu = n - k - 2$ . The student can verify that the conclusions will be the same, for the same level of significance, as were obtained when the more laborious  $F$  distribution was used.

In the present case  $r_{13.2}$  and  $b_{13.2}$  are not significant. The evidence that the finger-dexterity test has any value for estimating the criterion, in addition to that provided by the manual-dexterity test, is not convincing. Since  $r_{13.2}$  and  $b_{13.2}$  are not significant, we should test  $r_{12}$  and  $b_{12}$ . Appendix 10 shows that  $r_{12} = +0.648$  exceeds the rejection value of  $r_{12}$  at  $\alpha = 0.025$  and  $\nu = n - 2 = 25$ ; i.e.,  $0.648 > 0.381$ . Therefore,  $r_{12}$  and  $b_{12}$  are significant.

An alternative procedure is to test  $r_{12}$  and  $r_{13}$  for significance and then, if both are significant, to proceed with the computation of  $r_{12.3}$  and  $r_{13.2}$ . If both of these are significant, at some appropriate level, the multiple estimating equation is justified. In the present instance  $r_{13}$  is significant almost at the 0.001 level, but  $r_{13.2}$  is not significant even at the 0.10 level. The reason for this is the great amount of duplication between  $X_2$  and  $X_3$ . Chart 16.2 shows that the joint contribution of  $X_2$  and  $X_3$  to the explained variation is much larger than the net contribution of either of them and almost as large as their gross contribution.

## 16.8 TESTING OTHER HYPOTHESES CONCERNING PARTIAL CORRELATION

The  $z$  transformation can be used for the partial, but not multiple, correlation coefficient to test hypotheses and set confidence limits in the same manner as was described in Secs. 15.6 and 15.7 for the simple correlation coefficient. However, the standard error is different. In general,

$$\sigma_{r_{1m.23\dots(m-1)}} = \frac{1}{\sqrt{n-m-2/3}} \quad (16-21)$$

where  $m$  is the number of constants in the estimating equation. Otherwise, the procedure is the same as previously described, and the transformation is used to test the hypothesis that  $\rho_{12.3}$  is some value other than zero, or to test the significance of difference between two partial correlation coefficients.

## 16.9 TRANSFORMED DATA

In this section we will discuss some common methods of transformation which impart great flexibility to multiple regression analysis.

**Polynomial Regression.** An equation of the form

$$\hat{X}_1 = a + bX_2 + cX_2^2 + dX_2^3 + \dots$$

may be fit by using multiple regression analysis by letting  $X_3 = X_2^2$ ,  $X_4 = X_2^3$ , etc. Specific examples of polynomial regression are given in Chapter 19.

**Multiple Nonlinear Regression.** Suppose that we have one dependent variable  $Y$  and two independent variables  $W$  and  $X$  and a relationship of the type

$$Y = a + bW + cW^2 + dX + eX^2 + fX^3$$

Now let  $X_1 = Y$ ,  $X_2 = W$ ,  $X_3 = W^2$ ,  $X_4 = X$ ,  $X_5 = X^2$ , and  $X_6 = X^3$ . The estimating equation can then be written

$$\hat{X}_1 = a + b_{12.3456}X_2 + b_{13.2456}X_3 + \dots + b_{16.2345}X_6$$

The problem can be handled by a simple extension of the methods of the previous sections. If a chart similar to Chart 16.1 were drawn for the above relationship, with horizontal axes for  $W$  and  $X$  and a vertical axis for  $Y$ , all of the curves for vertical sections taken parallel to  $W$  and  $X$  would be nonlinear, but curves for parallel sections would be parallel.

**Regression with Interaction.** Sometimes the relationship between variables 1 and 2 may change as variable 3 changes. For example, the amount spent for some commodity may vary with age and income of the persons considered, and the effect of income on the amount spent may vary

with age. An appropriate type of equation might be

$$Y = a + bW + cX + dWX$$

since the effect of income and age may be partly separate additive effects and partly a joint multiplicative effect. Now let  $X_1 = Y$ ,  $X_2 = W$ ,  $X_3 = X$ ,  $X_4 = WX$  and write the equation in the usual multiple regression form. If a chart similar to Chart 16.1 were drawn for such a relationship, all of the slopes would be linear for vertical sections taken parallel to the  $X_2$  or  $X_3$  axis, but the steepness of the slope would depend on where the section was taken.

Sometimes relationships are both nonlinear and joint. For example,

$$Y = a + bW + cX + dW^2 + eWX + fX^2$$

This may be written in the usual form with  $X_1 = Y$ ,  $X_2 = W$ ,  $X_3 = X$ ,  $X_4 = W^2$ ,  $X_5 = WX$ , and  $X_6 = X^2$ .

## 16.10 THE MULTIPLE-PARTIAL CORRELATION COEFFICIENT

The multiple-partial correlation coefficient is the coefficient of multiple correlation of the dependent variable with two or more independent variables when all of the variables have been adjusted for the effect of one or more other variables.<sup>(12)</sup> For example:

$r_{1(234).5}$  is the coefficient of multiple correlation of  $x_{1.5}$  with  $x_{2.5}$ ,  $x_{3.5}$ , and  $x_{4.5}$ .

$r_{1(45).23}$  is the coefficient of multiple correlation of  $x_{1.23}$  with  $x_{4.23}$  and  $x_{5.23}$ .

In general, 
$$r_{i(jk\dots).tu\dots}^2 = \frac{\sum x_{i(jk\dots).tu\dots}^2}{\sum x_{i.tu\dots}^2} \quad (16-22)$$

Thus 
$$r_{1(345).2}^2 = \frac{\sum x_{1(345).2}^2}{\sum x_{1.2}^2}$$

where 
$$\sum x_{1(345).2}^2 = \sum x_{1(2345)}^2 - \sum x_{1(2)}^2$$

and 
$$\sum x_{1.2}^2 = \sum x_1^2 - \sum x_{1(2)}^2$$

Similarly, 
$$r_{1(45).23}^2 = \frac{\sum x_{1(45).23}^2}{\sum x_{1.23}^2}$$

where 
$$\sum x_{1(45).23}^2 = \sum x_{1(2345)}^2 - \sum x_{1(23)}^2$$

and 
$$\sum x_{1.23}^2 = \sum x_1^2 - \sum x_{1(23)}^2$$

<sup>(12)</sup> For an expository article on this coefficient, see Dudley J. Cowden, "The Multiple-Partial Correlation Coefficient," *Journal of the American Statistical Association*, Vol. 47 (September, 1952), pp. 442-456. A number of methods of computation are given in that article, including the method suggested in this text. Testing significance is also explained.

## 16.11 ADJUSTED COEFFICIENT OF DETERMINATION

A widely accepted scientific principle is that a simple theory is to be preferred to a complicated theory, other things equal. As variables are added to a multiple estimating equation, the postulated relationship becomes ever more involved. Therefore, it is desirable to formulate the estimating equation by using the fewest possible number of variables. A multiple coefficient of determination that places a penalty on the addition of another variable is sometimes called the *adjusted* coefficient of multiple determination and is given by<sup>(13)</sup>

$$\begin{aligned}\tilde{r}_{1(23\dots m)}^2 &= 1 - \frac{s_{1.23\dots m}^2}{s_1^2} \\ &= 1 - (1 - r_{1(23\dots m)}^2) \left( \frac{n-1}{n-m} \right)\end{aligned}$$

Thus, the adjusted coefficient eventually decreases as  $m$  increases, given a fixed sample size. This adjusted coefficient may also be thought of as an estimator for  $\rho_{1(23\dots m)}^2$ .

Multiple regression analysis has been widely programmed for electronic computers. Sometimes one wishes to select the "best" set of independent variables from a larger list of candidates in terms of their power to predict the dependent variable. One approach would be to program a computer to regress  $X_1$  on every pair, triplet, etc., of  $m'$  candidates. The combination of  $m$  variables giving the largest adjusted coefficient of determination might then be selected. For a study such as this a computer is almost absolutely essential, since  $2^{m'-1}$  regression equations must be computed. Even then, when  $m'$  is only modestly large calculation becomes prohibitively expensive. For example, if  $m' = 24$ , then 16,777,215 equations must be calculated.

To reduce the expense of calculation, *stepwise* regression routines have been developed. One type of stepwise routine brings into the regression equation at each step the independent variable with the highest partial correlation with the dependent variable, considering the variables already in the equation. The process ends when no variable can be added that will give a criterion statistic (such as an  $F$  statistic) larger than some predetermined (often arbitrary) level. Variables once entered into the equation can be dropped if the addition of a new variable causes the criterion statistic of the variable to be removed to fall below some predetermined level.<sup>(14)</sup> In some cases, variables once dropped from the equation may be reentered into the equation.

<sup>(13)</sup> The student can derive this coefficient by following the argument given for Eq. (A15-8). Also, the symbol  $\bar{R}$  is often found in the literature.

<sup>(14)</sup> For an example program for the stepwise technique, see: W. J. Dixon, ed., *BMD, Biomedical Computer Programs*, (Berkeley and Los Angeles: University of California Press, 1967), pp. 233-257d.



Stepwise regression should be carried out with caution. The technique is tantamount to fishing for significantly correlated variables, and tests of significance based upon classical statistical methods are invalidated since the hypothesis itself was developed by regression. Thus if a regression equation has been formulated by a stepwise technique, it should be recomputed from another independent sample before its coefficients are tested for significance. As a final word of caution, different types of stepwise programs will not always select the same independent variables from a list of candidates, nor will a given routine necessarily select the optimum set of candidates in terms of the adjusted coefficient of determination. The researcher who uses a stepwise computer program or any prewritten computer program should have a complete understanding of what the program does and does not accomplish.

## PROBLEMS

1. Show that

$$r_{13.2}^2 = \frac{r_{1(23)}^2 - r_{12}^2}{1 - r_{12}^2}$$

2. Verify the equality of:

- a. Equations (16-14) and (16-15).  
b. Equations (16-16) and (16-17).

Also show that  $t_{12.3} = \sqrt{F_{12.3}}$  as they are given in Eqs. (16-20) and (16-17).

3. Beta weights:

- a. Show that

$$\beta_{12} = b_{12} \sqrt{\frac{\sum x_2^2}{\sum x_1^2}}$$

- b. Let  $z_1 = (X_1 - \bar{X}_1)/s_1$  and  $z_2 = (X_2 - \bar{X}_2)/s_2$ . Now show that the regression equation calculated from  $z_1$  and  $z_2$  is  $z_{1(2)} = \beta_{12}z_2$ . Extend the argument to three  $z$  variables.

- c. Argue that  $\beta_{12}$  tells us that with one standard deviation change in  $X_2$  the estimated change in  $X_1$  is  $\beta_{12}s_1$ . Extend this argument to  $\beta_{12.3}$ .

- d. Calculate  $\beta_{12}$  and  $\beta_{12.3}$ , using the example data given in this chapter.

- e. Show that  $r_{12.3}^2 = (\beta_{12.3})(\beta_{21.3})$  and that  $r_{12} = \beta_{12}$ .

4. Given the following matrix of simple (zero order) correlation coefficients, find the second order coefficient  $r_{13.24}$ . How many first and second order coefficients are there?

$$\begin{bmatrix} r_{11} & r_{12} & r_{13} & r_{14} \\ & r_{22} & r_{23} & r_{24} \\ & & r_{33} & r_{34} \\ & & & r_{44} \end{bmatrix} = \begin{bmatrix} 1.0 & 0.4491 & 0.3869 & 0.3307 \\ & 1.0 & 0.7992 & 0.6690 \\ & & 1.0 & 0.7789 \\ & & & 1.0 \end{bmatrix}$$

Assuming  $n = 25$  and  $m = 4$ , perform a two-sided test of the significance of  $r_{13,24}$  using  $F$ ,  $t$ , and Appendix 10. Let  $\alpha = 0.05$ .

5. Show that just as

$$r_{12} = b_{12} \frac{s_2}{s_1}$$

so

$$r_{12.3} = b_{12.3} \frac{s_{2.13}}{s_{1.23}}$$

and

$$r_{12.34} = b_{12.34} \frac{s_{2.134}}{s_{1.234}}$$

## APPENDIX: Matrix Algebra Approach to Regression

Much of the literature concerning multiple regression analysis is couched in terms of matrix algebra, since the notation system is quite convenient. Therefore, we restate much of the material in Chapters 14 through 16 in terms of matrix algebra. We will also prove some of the assertions of these chapters. Specifically, the discussion will use the three-variable multiple regression example of Chapter 16, but the techniques may be easily extended to an arbitrary number of variables.

### A16.1 SOME ELEMENTS OF MATRIX ALGEBRA

A matrix is an ordered set of elements (numbers) that are arranged in rows and columns. A matrix that has  $r$  rows and  $c$  columns is said to be of order  $r \times c$  ( $r$  by  $c$ ). Symbolically, if  $\mathbf{A}$  is an  $r \times c$  matrix,

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1c} \\ a_{21} & a_{22} & \cdots & a_{2c} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{r1} & a_{r2} & \cdots & a_{rc} \end{bmatrix}$$

Letters in bold type, such as  $\mathbf{A}$  will be used to symbolize a matrix. The elements of the matrix,  $a_{11}$ ,  $a_{12}$  . . . etc., will be enclosed by brackets. The subscripts on the elements will, in this section, refer in order to the row number and column number of the matrix where the element is located. For example  $a_{23}$  is the element in the second row and third column of  $\mathbf{A}$ .

Mathematical operations with matrices do not necessarily follow the rules of ordinary algebra. Below are illustrated some selected mathematical operations with matrices.

1. *Addition and Subtraction.* Two matrices may be added together or subtracted from each other only if they are both of the same order. Addition is carried out by adding together the corresponding elements of the two matrices

RULE		EXAMPLE		
$A + B = B + A$				$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} + \begin{bmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{bmatrix} = \begin{bmatrix} 2 & 6 & 10 \\ 6 & 10 & 14 \\ 10 & 14 & 18 \end{bmatrix}$

Subtraction is the reverse of addition

RULE		EXAMPLE		
$A - B = A + (-B)$				$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} - \begin{bmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{bmatrix} = \begin{bmatrix} 0 & -2 & -4 \\ 2 & 0 & -2 \\ 4 & 2 & 0 \end{bmatrix}$

2. *Multiplication.* Multiplication of matrices does not follow ordinary rules of algebra, except in certain restrictive instances. The simplest type of multiplication is the multiplication of a matrix by a single real number called a *scalar*. In this case, each element of the matrix is simply multiplied by the scalar. For example, if the scalar is denoted as  $d$ ,

RULE		
$d \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1c} \\ a_{21} & a_{22} & \cdots & a_{2c} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{r1} & a_{r2} & \cdots & a_{rc} \end{bmatrix} = \begin{bmatrix} (d)a_{11} & (d)a_{12} & \cdots & (d)a_{1c} \\ (d)a_{21} & (d)a_{22} & \cdots & (d)a_{2c} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ (d)a_{r1} & (d)a_{r2} & \cdots & (d)a_{rc} \end{bmatrix}$		

EXAMPLE	
	$2 \begin{bmatrix} 1 & -2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix} = \begin{bmatrix} 2 & -4 \\ 6 & 8 \\ 10 & 12 \end{bmatrix}$

Two matrices may be multiplied together only if the matrices are *conformable* under the desired order of multiplication. Two matrices are conformable if the number of columns in one matrix is the same as the number of rows in the other. The product matrix  $P$  may be defined as

$$P = AB$$

where **A** and **B** are matrices. The  $ij$ th element of **P** is found by multiplying each element in the  $i$ th row of matrix **A** by each element in the  $j$ th column of matrix **B** and summing the resulting terms. It is clear that the number of columns in matrix **A** must be the same as the number of rows in matrix **B** to carry out this operation. For example, if

$$\mathbf{A} = \begin{bmatrix} 2 & 1 & 1 \\ 3 & 2 & 4 \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} 2 & 1 \\ 3 & 2 \\ 1 & 1 \end{bmatrix}$$

we see that **A** is of order  $2 \times 3$  and **B** is of order  $3 \times 2$ . Since the number of columns in matrix **A** is the same as the number of rows in matrix **B**, the product **P** = **AB** is possible and is

$$\mathbf{P} = \begin{bmatrix} 2(2) + 1(3) + 1(1) & 2(1) + 1(2) + 1(1) \\ 3(2) + 2(3) + 4(1) & 3(1) + 2(2) + 4(1) \end{bmatrix}$$

or 
$$\mathbf{P} = \begin{bmatrix} 8 & 5 \\ 16 & 11 \end{bmatrix}$$

Notice that although the product **AB** may be possible, it does not follow that the product **BA** will be possible, as we will show below. Also, even if both products are possible, it does *not* follow that **BA** = **AB**. Hence, in matrix multiplication the order of multiplication is important. The student may verify that

$$\mathbf{BA} = \begin{bmatrix} 7 & 4 & 6 \\ 12 & 7 & 11 \\ 5 & 3 & 5 \end{bmatrix}$$

The following device is extremely helpful in remembering whether or not two matrices may be multiplied together at all, whether this multiplication may take place for both **AB** and **BA**, and what the order of the resulting product matrix will be. First, write down the order of the two matrices. For example,

$$(2 \times 3) \quad \text{and} \quad (3 \times 2)$$

If the two inside numbers match, the matrices are conformable under the order of multiplication desired, since this will indicate that the number of columns in the first matrix is the same as the number of rows in the second matrix. The two outside numbers will give the order of the resulting product matrix. For example, if **A** is  $3 \times 3$  and **B** is  $3 \times 6$ , then

$$(3 \times 3) \quad \text{and} \quad (3 \times 6)$$

shows that the multiplication **AB** is possible, since the inside numbers match, and the outside numbers show that the resulting product matrix will

be of the order  $3 \times 6$ . However, the multiplication  $\mathbf{BA}$  is not possible, since

$$(3 \times 6) \text{ and } (3 \times 3)$$

do not have matching inside numbers.

3. *Transposition.* The transpose of the matrix  $\mathbf{A}$ , denoted by  $\mathbf{A}'$ , is accomplished by interchanging the rows and columns of the matrix. For example, if

$$\mathbf{A} = \begin{bmatrix} 2 & 3 \\ 1 & 4 \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} 1 & 5 \\ 2 & 3 \\ 4 & 6 \end{bmatrix}$$

then 
$$\mathbf{A}' = \begin{bmatrix} 2 & 1 \\ 3 & 4 \end{bmatrix} \quad \text{and} \quad \mathbf{B}' = \begin{bmatrix} 1 & 2 & 4 \\ 5 & 3 & 6 \end{bmatrix}$$

Notice that the transpose of the transposed matrix is the original matrix.

RULE	EXAMPLE
$(\mathbf{A}')' = \mathbf{A}$	$\begin{bmatrix} 2 & 1 \\ 3 & 4 \end{bmatrix}' = \begin{bmatrix} 2 & 3 \\ 1 & 4 \end{bmatrix}$

Also notice that transposing a *symmetric* matrix has no effect on the matrix at all. A symmetric matrix is one whose elements above the main, or north-west-southeast, diagonal are the mirror image of the elements below the main diagonal. The following matrix is symmetric.

$$\mathbf{A} = \begin{bmatrix} 2 & 1 & 5 \\ 1 & 3 & 6 \\ 5 & 6 & 4 \end{bmatrix}$$

and the student may verify that  $\mathbf{A}' = \mathbf{A}$ .

4. *Inversion.* In ordinary algebra if  $b$  is a nonzero number

$$\frac{b}{b} = bb^{-1} = 1$$

In matrix algebra the process of division is not defined, but analogous to the concept of division in ordinary algebra is the concept of the *inverse matrix* in matrix algebra. The inverse of the matrix  $\mathbf{A}$ , if it exists, is defined as the matrix  $\mathbf{A}^{-1}$ , such that

$$\mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$$

where  $\mathbf{I}$  is the *unit matrix* of the same order as the matrix  $\mathbf{A}$ .

The unit matrix is a square matrix that consists only of the numbers 1 on its main diagonal and zeros everywhere else. The unit matrix is, therefore,

symmetric. For example a  $3 \times 3$  unit matrix is

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Since the unit matrix must be square, it follows that for a matrix to have an inverse it must be square. For example, if

$$\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 0 & 2 \end{bmatrix}$$

the inverse of  $\mathbf{A}$  is

$$\mathbf{A}^{-1} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{4} \\ 0 & \frac{1}{2} \end{bmatrix}$$

and we see that

$$\begin{bmatrix} 2 & 1 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & -\frac{1}{4} \\ 0 & \frac{1}{2} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{4} \\ 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

There are many ways of finding the inverse of a matrix when it exists.<sup>(1)</sup> We will illustrate one such method.

For any square matrix  $\mathbf{A}$  there exists a scalar which is called the *determinant* of  $\mathbf{A}$ . We will not give a rigorous mathematical definition of a determinant but rather will illustrate two orders of determinants. The student should already be familiar with the notion of a determinant. For a  $2 \times 2$  matrix, the determinant will be the product of the elements on the main diagonal minus the product of the elements on the southwest-northeast diagonal (secondary diagonal). Thus if

$$\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 0 & 2 \end{bmatrix}$$

the determinant of  $\mathbf{A}$  is

$$|\mathbf{A}| = \begin{vmatrix} 2 & 1 \\ 0 & 2 \end{vmatrix} = 2(2) - 0(1) = 4$$

We will use vertical lines around a matrix to indicate the determinant of the matrix. The determinant of an  $n \times n$  matrix will be called a determinant of the  $n$ th order. Thus directly above we have a second order determinant. Higher order determinants may be found fairly easily by the evaluation of the lower order determinants, called *minors*. For example, to evaluate the determinant of the matrix

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

<sup>(1)</sup> See, for example, V. N. Faddeeva, *Computational Methods in Linear Algebra* (New York: Dover Publications, Inc., 1959).

we may form the expansion<sup>(2)</sup>

$$|A| = a_{11} \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{21} \begin{vmatrix} a_{12} & a_{13} \\ a_{32} & a_{33} \end{vmatrix} + a_{31} \begin{vmatrix} a_{12} & a_{13} \\ a_{22} & a_{23} \end{vmatrix}$$

Notice that the expansion was formed by taking the elements in the first column and writing them down, i.e.,  $a_{11}$ ,  $a_{21}$ , and  $a_{31}$ . Let us call these elements expansion elements. For each expansion element the row and column of the original matrix from which that expansion element was drawn are mentally crossed out, and the remainder of the matrix is written down to the right of the expansion element as a *minor determinant*, or simply *minor*. The resulting terms are then summed according to a certain sign scheme. The sign scheme for summation may be determined by adding the subscripts on the expansion element. If the sum of the subscripts is an even number, the minor is multiplied by  $+1$ ; if the sum is an odd number, the minor is multiplied by  $-1$ . Thus

$$|A| = a_{11}(-1)^{1+1} \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} + a_{21}(-1)^{2+1} \begin{vmatrix} a_{12} & a_{13} \\ a_{32} & a_{33} \end{vmatrix} + a_{31}(-1)^{3+1} \begin{vmatrix} a_{12} & a_{13} \\ a_{22} & a_{23} \end{vmatrix}$$

since  $-1$  raised to any odd power is  $-1$ , and  $-1$  raised to any even power is  $+1$ . Notice that the signs alternate, so they may be written down directly after determining the first sign.

To give a numerical example of the process of evaluating a determinant of the third order, consider the following matrix.

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

Then

$$|A| = 1 \begin{vmatrix} 5 & 6 \\ 8 & 9 \end{vmatrix} - 4 \begin{vmatrix} 2 & 3 \\ 8 & 9 \end{vmatrix} + 7 \begin{vmatrix} 2 & 3 \\ 5 & 6 \end{vmatrix}$$

$$\text{or } |A| = 1(45 - 48) - 4(18 - 24) + 7(12 - 15) = -3 + 24 - 21 = 0$$

Having illustrated the concept of a determinant, we will now show its usefulness in inverting matrices. To find the inverse of a matrix, the following five steps may be followed:

1. *Evaluate the determinant of the matrix.* If the determinant is zero, as in the illustration above, the matrix has no inverse and is called a *singular* matrix. If the determinant is nonzero, the matrix has an inverse and is called *nonsingular*. For example, if

$$A = \begin{bmatrix} 1 & 3 & 2 \\ 0 & 5 & 1 \\ 2 & 3 & 2 \end{bmatrix}$$

<sup>(2)</sup> It should be noted that the expansion by minors need not necessarily be done about the first column of a matrix. Any row or column may be used, presumably the one with the most zeros to facilitate computation.

then

$$|A| = 1 \begin{vmatrix} 5 & 1 \\ 3 & 2 \end{vmatrix} - 0 \begin{vmatrix} 3 & 2 \\ 3 & 2 \end{vmatrix} + 2 \begin{vmatrix} 3 & 2 \\ 5 & 1 \end{vmatrix}$$

$$|A| = 1(7) - 0(0) + 2(-7) = -7$$

Since the determinant is nonzero, we may proceed to the next step.

2. *Establish a matrix of cofactors from the original matrix.* The matrix of cofactors is a matrix of signed minor determinants that are associated with each of the elements in the original matrix. In the example directly below, the entry in the extreme northwest corner was found by replacing the element 1 in the original matrix by the determinant found after eliminating the elements in the first row and column of the original matrix. The second entry in the first row was found by replacing the element 3 by the determinant found after eliminating the first row and second column of the original matrix. Notice also that this determinant is preceded by a minus sign in accordance with the sign scheme for minor determinants previously discussed.

$$\begin{bmatrix} \begin{vmatrix} 5 & 1 \\ 3 & 2 \end{vmatrix} & -\begin{vmatrix} 0 & 1 \\ 2 & 2 \end{vmatrix} & \begin{vmatrix} 0 & 5 \\ 2 & 3 \end{vmatrix} \\ -\begin{vmatrix} 3 & 2 \\ 3 & 2 \end{vmatrix} & \begin{vmatrix} 1 & 2 \\ 2 & 2 \end{vmatrix} & -\begin{vmatrix} 1 & 3 \\ 2 & 3 \end{vmatrix} \\ \begin{vmatrix} 3 & 2 \\ 5 & 1 \end{vmatrix} & -\begin{vmatrix} 1 & 2 \\ 0 & 1 \end{vmatrix} & \begin{vmatrix} 1 & 3 \\ 0 & 5 \end{vmatrix} \end{bmatrix} = \begin{bmatrix} 7 & 2 & -10 \\ 0 & -2 & 3 \\ -7 & -1 & 5 \end{bmatrix}$$

3. *Form the adjoint of the original matrix, denoted as adj A.* The adjoint of a matrix is the transpose of the matrix of cofactors. Thus, for our example,

$$\text{adj } A = \begin{bmatrix} 7 & 0 & -7 \\ 2 & -2 & -1 \\ -10 & 3 & 5 \end{bmatrix}$$

4. *Evaluate the inverse of the matrix A.* The inverse of the matrix is now found by multiplying each element of the adjoint matrix by the reciprocal of the determinant of the original matrix, i.e., by  $1/|A|$ .

Thus

$$A^{-1} = \frac{1}{|A|} (\text{adj } A)$$

It should be clear now why a matrix that possesses a determinant of zero cannot be inverted, since if  $|A| = 0$ , then  $1/|A|$  is undefined. The inverse of matrix A is now found to be

$$A^{-1} = -\frac{1}{7} \begin{bmatrix} 7 & 0 & -7 \\ 2 & -2 & -1 \\ -10 & 3 & 5 \end{bmatrix} = \begin{bmatrix} -\frac{7}{7} & \frac{0}{7} & \frac{7}{7} \\ -\frac{2}{7} & \frac{2}{7} & \frac{1}{7} \\ \frac{10}{7} & -\frac{3}{7} & -\frac{5}{7} \end{bmatrix} = \begin{bmatrix} -1 & 0 & 1 \\ -\frac{2}{7} & \frac{2}{7} & \frac{1}{7} \\ \frac{10}{7} & -\frac{3}{7} & -\frac{5}{7} \end{bmatrix}$$



5. *Check the inverse.* We see that

$$\begin{bmatrix} 1 & 3 & 2 \\ 0 & 5 & 1 \\ 2 & 3 & 2 \end{bmatrix} \begin{bmatrix} -1 & 0 & 1 \\ -\frac{2}{7} & \frac{2}{7} & \frac{1}{7} \\ \frac{10}{7} & -\frac{3}{7} & -\frac{5}{7} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Matrix inversion is particularly important in the solution of certain systems of simultaneous linear equations. Consider the following three equations:

$$X_1 + 3X_2 + 2X_3 = 13$$

$$5X_2 + X_3 = 13$$

$$2X_1 + 3X_2 + 2X_3 = 14$$

These equations may be written in matrix form as

$$\begin{bmatrix} 1 & 3 & 2 \\ 0 & 5 & 1 \\ 2 & 3 & 2 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} 13 \\ 13 \\ 14 \end{bmatrix}$$

i.e., in the form

$$\mathbf{BX} = \mathbf{C}$$

where **B** represents the coefficients of the **X** variables and **C**, the matrix of constant terms to the right of the equality sign. Notice that both **X** and **C** are single column matrices, which are often called *column vectors*. A row vector is the transpose of a column vector.

In ordinary algebra if we had the single equation

$$bX = c$$

we could solve for *X* by evaluating

$$X = \frac{c}{b} = b^{-1}c$$

if *b* and *c* are constants and *b* is not zero. In matrix algebra we solve for the column vector **X** by evaluating

$$\mathbf{X} = (\mathbf{B}^{-1})\mathbf{C}$$

if **B** is nonsingular. Thus, drawing on previous results, we find

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} -1 & 0 & 1 \\ -\frac{2}{7} & \frac{2}{7} & \frac{1}{7} \\ \frac{10}{7} & -\frac{3}{7} & -\frac{5}{7} \end{bmatrix} \begin{bmatrix} 13 \\ 13 \\ 14 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

It is clear that the solutions to the simultaneous equations are  $X_1 = 1$ ,  $X_2 = 2$ , and  $X_3 = 3$ .

## A16.2 APPLICATION OF MATRIX ALGEBRA TO MULTIPLE REGRESSION ANALYSIS

Let the observations on a dependent variable  $X_1$  and the observations on two independent variables  $X_2$  and  $X_3$  be related in the following way:

$$X_1 = a + b_{12.3}X_2 + b_{13.2}X_3 + e \quad (\text{A16-1})$$

where  $e$  is the "error term" or simply the deviation in  $X_1$  not explained by  $X_2$  and  $X_3$ . Thus

$$e = X_1 - \hat{X}_1$$

so that Eq. (A16-1) may be written as Eq. (16-1) in Sec. 16.1.

$$\hat{X}_1 = a + b_{12.3}X_2 + b_{13.2}X_3 \quad (\text{A16-1a})$$

Let us define the four matrices

$$\mathbf{Y} = \begin{bmatrix} X_{11} \\ X_{12} \\ \cdot \\ \cdot \\ X_{1n} \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & X_{21} & X_{31} \\ 1 & X_{22} & X_{32} \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 1 & X_{2n} & X_{3n} \end{bmatrix}$$

$$\mathbf{B} = \begin{bmatrix} a \\ b_{12.3} \\ b_{13.2} \end{bmatrix} \quad \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \cdot \\ \cdot \\ e_n \end{bmatrix}$$

Notice in the  $\mathbf{Y}$  and  $\mathbf{X}$  matrices that the first subscript refers to the variable in question, and the second subscript refers to the observation on that variable rather than to the row and column number of the matrix.

From Eqs. (16-3) we know that the normal equations needed to evaluate the coefficients of Eq. (A16-1) are found by minimizing

$$\sum e^2 = \sum (X_1 - \hat{X}_1)^2$$

and are

$$\begin{aligned} na + b_{12.3} \sum X_2 + b_{13.2} \sum X_3 &= \sum X_1 & \text{I} \\ a \sum X_2 + b_{12.3} \sum X_2^2 + b_{13.2} \sum X_2 X_3 &= \sum X_1 X_2 & \text{II} \\ a \sum X_3 + b_{12.3} \sum X_2 X_3 + b_{13.2} \sum X_3^2 &= \sum X_1 X_3 & \text{III} \end{aligned}$$

In terms of matrix algebra we may write these normal equations:

$$\begin{bmatrix} n & \sum X_2 & \sum X_3 \\ \sum X_2 & \sum X_2^2 & \sum X_2 X_3 \\ \sum X_3 & \sum X_2 X_3 & \sum X_3^2 \end{bmatrix} \begin{bmatrix} a \\ b_{12.3} \\ b_{13.2} \end{bmatrix} = \begin{bmatrix} \sum X_1 \\ \sum X_1 X_2 \\ \sum X_1 X_3 \end{bmatrix}$$

and we notice that

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_{21} & X_{22} & \cdots & X_{2n} \\ X_{31} & X_{32} & \cdots & X_{3n} \end{bmatrix} \begin{bmatrix} 1 & X_{21} & X_{31} \\ 1 & X_{22} & X_{32} \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 1 & X_{2n} & X_{3n} \end{bmatrix} = \begin{bmatrix} n & \Sigma X_2 & \Sigma X_3 \\ \Sigma X_2 & \Sigma X_2^2 & \Sigma X_2 X_3 \\ \Sigma X_3 & \Sigma X_2 X_3 & \Sigma X_3^2 \end{bmatrix}$$

and

$$\mathbf{X}'\mathbf{Y} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_{21} & X_{22} & \cdots & X_{2n} \\ X_{31} & X_{32} & \cdots & X_{3n} \end{bmatrix} \begin{bmatrix} X_{11} \\ X_{12} \\ \cdot \\ \cdot \\ X_{1n} \end{bmatrix} = \begin{bmatrix} \Sigma X_1 \\ \Sigma X_1 X_2 \\ \Sigma X_1 X_3 \end{bmatrix}$$

so that the normal equations can be written compactly as

$$(\mathbf{X}'\mathbf{X})\mathbf{B} = \mathbf{X}'\mathbf{Y} \quad (\text{A16-2})$$

and Eq. (A16-1) can be written as

$$\mathbf{Y} = \mathbf{XB} + \mathbf{e}$$

for all observations on the variables.

The solution to the normal equations given by Eq. (A16-2) is

$$\mathbf{B} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

provided that  $\mathbf{X}'\mathbf{X}$  is nonsingular. To illustrate, we find from Table 16.1 and Table 16.2

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 27 & 3,190 & 2,269 \\ 3,190 & 380,040 & 269,820 \\ 2,269 & 269,820 & 193,021 \end{bmatrix}$$

and

$$\mathbf{X}'\mathbf{Y} = \begin{bmatrix} 4,654 \\ 556,637 \\ 396,486 \end{bmatrix}$$

The inverse of the  $\mathbf{X}'\mathbf{X}$  matrix is approximately

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 4.724735663 & -0.030042200 & -0.013544844 \\ -0.030042200 & 0.000540150 & -0.000401912 \\ -0.013544844 & -0.000401912 & 0.000726227 \end{bmatrix}$$

Then

$$\begin{bmatrix} a \\ b_{12.3} \\ b_{13.2} \end{bmatrix} = \begin{bmatrix} 4.724735663 & -0.030042200 & -0.013544844 \\ -0.030042200 & 0.000540150 & -0.000401912 \\ -0.013544844 & -0.000401912 & 0.000726227 \end{bmatrix} \begin{bmatrix} 4,654 \\ 556,637 \\ 396,486 \end{bmatrix}$$

$$= \begin{bmatrix} -104.02 \\ 1.498 \\ 1.182 \end{bmatrix}$$

which is the same result obtained in Sec. 16.3. Notice that both  $\mathbf{X}'\mathbf{X}$  and  $(\mathbf{X}'\mathbf{X})^{-1}$  are symmetric.

The order of the matrix to be inverted can be reduced by one if the observations are expressed as deviations from their respective means.

Thus 
$$x_1 = b_{12.3}x_2 + b_{13.2}x_3 + e \quad (\text{A16-3})$$

where<sup>(3)</sup> 
$$e = x_1 - x_{1(23)}$$

We may write Eq. (A16-3) as Eq. (16-4).

$$x_{1(23)} = b_{12.3}x_2 + b_{13.2}x_3 \quad (\text{A16-3a})$$

If we define the three matrices

$$\mathbf{x} = \begin{bmatrix} x_{21} & x_{31} \\ x_{22} & x_{32} \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ x_{2n} & x_{3n} \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} x_{11} \\ x_{12} \\ \cdot \\ \cdot \\ \cdot \\ x_{1n} \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} b_{12.3} \\ b_{13.2} \end{bmatrix}$$

and notice that

$$\mathbf{x}'\mathbf{x} = \begin{bmatrix} \sum x_2^2 & \sum x_2x_3 \\ \sum x_2x_3 & \sum x_3^2 \end{bmatrix} \quad \text{and} \quad \mathbf{x}'\mathbf{y} = \begin{bmatrix} \sum x_1x_2 \\ \sum x_1x_3 \end{bmatrix}$$

we see that Eq. (A16-3) can be expressed as

$$\mathbf{y} = \mathbf{xb} + \mathbf{e}$$

for all observations on the variables. Also the normal equations necessary to estimate  $\mathbf{b}$  are

$$(\mathbf{x}'\mathbf{x})\mathbf{b} = \mathbf{x}'\mathbf{y}$$

From Table 16.2 we see that

$$\mathbf{x}'\mathbf{x} = \begin{bmatrix} 3147.41 & 1741.85 \\ 1741.85 & 2340.96 \end{bmatrix} \quad \text{and} \quad \mathbf{x}'\mathbf{y} = \begin{bmatrix} 6775.52 \\ 5377.62 \end{bmatrix}$$

Then

$$\mathbf{b} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y}$$

<sup>(3)</sup> Notice that the meaning of  $e$  is unchanged from the way it was defined for Eq. (A16-1).

$$e = x_1 - x_{1(23)} = (X_1 - \bar{X}_1) - (\hat{X}_1 - \bar{X}_1) = X_1 - \hat{X}_1$$

or, after inverting  $\mathbf{x}'\mathbf{x}$  we have

$$\begin{bmatrix} b_{12.3} \\ b_{13.2} \end{bmatrix} = \begin{bmatrix} 0.000540150 & -0.000401912 \\ -0.000401912 & 0.000726227 \end{bmatrix} \begin{bmatrix} 6775.52 \\ 5377.62 \end{bmatrix} = \begin{bmatrix} 1.498 \\ 1.182 \end{bmatrix}$$

The intercept can be found by evaluating

$$a = \bar{X}_1 - b_{12.3}\bar{X}_2 - b_{13.2}\bar{X}_3$$

in the same way as was noted in Sec. 16.2. Notice that  $(\mathbf{x}'\mathbf{x})^{-1}$  is a submatrix of  $(\mathbf{X}'\mathbf{X})^{-1}$ .

Although the ability to solve the normal equations for  $a$ ,  $b_{12.3}$ , and  $b_{13.2}$  depends only upon the nonsingularity of the  $\mathbf{X}'\mathbf{X}$  matrix, the derivation of standard errors of these statistics requires stronger assumptions. Let us call the population regression equation

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

and specify that:

1. The expected value of  $\boldsymbol{\epsilon}$ , the population error term, is zero.

$$E(\boldsymbol{\epsilon}) = 0$$

2. The independent variables  $\mathbf{X}$  are fixed numbers not subject to statistical error.

3. For each of the  $n$  observations on the independent variables  $\mathbf{X}$ , there is a distribution of errors. Each of these distributions has the same variance  $\sigma^2$ , and successive errors are uncorrelated. These assumptions may be stated compactly as

$$\begin{aligned} E\{[\boldsymbol{\epsilon} - E(\boldsymbol{\epsilon})][\boldsymbol{\epsilon} - E(\boldsymbol{\epsilon})]'\} &= E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \sigma^2\mathbf{I} \\ &= \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} \end{aligned}$$

Now since

$$\mathbf{B} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

and

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

we find upon substitution that

$$\begin{aligned} \mathbf{B} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon} \end{aligned} \tag{A16-4}$$

since  $(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X}) = \mathbf{I}$ . Taking the expected value of both sides of the above equation, we see that since the  $\mathbf{X}$  matrix is a statistical constant,

$$E(\mathbf{B}) = \boldsymbol{\beta}$$

under the assumption that  $E(\epsilon) = 0$ . In other words,  $\mathbf{B}$  is an unbiased estimator of  $\beta$ .

The variance of the estimates can be found by evaluating

$$E\{[\mathbf{B} - E(\mathbf{B})][\mathbf{B} - E(\mathbf{B})]'\} = E[(\mathbf{B} - \beta)(\mathbf{B} - \beta)']$$

and using Eq. (A16-4)

$$\mathbf{B} - \beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon$$

and the fact that  $E(\mathbf{B}) = \beta$  results in our having

$$E[(\mathbf{B} - \beta)(\mathbf{B} - \beta)'] = E\{[\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon][\epsilon'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}]\}$$

This result follows because

$$[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon]' = [\epsilon'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}]$$

by virtue of the fact that  $(\mathbf{X}'\mathbf{X})^{-1}$  is symmetric and, therefore, unchanged by transposition, and by virtue of the fact that  $(\mathbf{X}')' = \mathbf{X}$ . Upon further simplification we find that

$$E[(\mathbf{B} - \beta)(\mathbf{B} - \beta)'] = \sigma^2\mathbf{I}(\mathbf{X}'\mathbf{X})^{-1}$$

since  $E(\epsilon\epsilon') = \sigma^2\mathbf{I}$ .

The variances of the statistics  $a$ ,  $b_{12.3}$ , and  $b_{13.2}$  may be found by multiplying the main diagonal of the  $(\mathbf{X}'\mathbf{X})^{-1}$  matrix by the scalar  $\sigma^2$ . However,  $\sigma^2$  is usually unknown. Fortunately, it can be estimated by use of

$$s_{1.23}^2 = \frac{\sum x_{1.23}^2}{n-3} = \frac{\mathbf{e}'\mathbf{e}}{n-3}$$

in the three-variable case, or generally by

$$s_{1.23\dots m}^2 = \frac{\sum x_{1.23\dots m}^2}{n-m} = \frac{\mathbf{e}'\mathbf{e}}{n-m}$$

To return to the numerical example of Chapter 16, we recall from Eq. (16-7) that we calculated

$$s_{1.23}^2 = 758.4$$

Therefore, using the main diagonal of  $(\mathbf{X}'\mathbf{X})^{-1}$  the variances of  $a$ ,  $b_{12.3}$ , and  $b_{13.2}$  are estimated to be

$$\begin{bmatrix} s_a^2 \\ s_{b_{12.3}}^2 \\ s_{b_{13.2}}^2 \end{bmatrix} = 758.4 \begin{bmatrix} 4.724735663 \\ 0.000540150 \\ 0.000726227 \end{bmatrix} = \begin{bmatrix} 3583.2 \\ 0.4096 \\ 0.5508 \end{bmatrix}$$

The estimated standard errors are the square root of these variances.

If we make the further assumption that the error term in the population is normally distributed, the following ratios have Student's distribution with  $n - m$  degrees of freedom.

$$t_a = \frac{a - \alpha}{s_a}, \quad t_{12.3} = \frac{b_{12.3} - \beta_{12.3}}{s_{b_{12.3}}}, \quad t_{13.2} = \frac{b_{13.2} - \beta_{13.2}}{s_{b_{13.2}}}$$

where  $\alpha$  is the population intercept. The use of these types of  $t$  statistics was illustrated in Sec. 15.5 in the case of simple regression. The same uses apply in multiple regression analysis.

First order partial correlation coefficients may be derived directly from the  $t$  statistics given above. Thus, under the hypothesis that the population parameters  $\rho_{12.3}$  and  $\rho_{13.2}$  are zero,

$$r_{12.3} = \frac{t_{12.3}}{\sqrt{(t_{12.3})^2 + (n - m)}}$$

$$r_{13.2} = \frac{t_{13.2}}{\sqrt{(t_{13.2})^2 + (n - m)}}$$

which can be deduced by solving Eqs. (16-20) for the partial correlation coefficients. For our numerical illustration

$$t_{12.3} = \frac{1.498}{0.640} = 2.341$$

$$\text{and } r_{12.3} = \frac{2.341}{\sqrt{(2.341)^2 + (27 - 3)}} = \frac{2.341}{5.43} = +0.431$$

and so on for  $r_{13.2}$ . The student can verify that the multiple correlation coefficient may be found by evaluating the square of

$$r_{1(23)}^2 = \frac{\sum x_1^2(23)}{\sum x_1^2} = \frac{\mathbf{b}'(\mathbf{x}'\mathbf{y})}{(\mathbf{y}'\mathbf{y})}$$

and hence that the elements needed for an  $F$  table such as that of Table 16.3 or Table 16.5 can be readily deduced.

### A16.3 THE DOOLITTLE COMPUTATION TECHNIQUE

Again, drawing on the three variable illustration, if we express all of the observations on all the variables in terms of deviations from their respective means, we may write the matrix<sup>(4)</sup>

$$\mathbf{X} = \begin{bmatrix} x_{21} & x_{31} & x_{11} \\ x_{22} & x_{32} & x_{12} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ x_{2n} & x_{3n} & x_{1n} \end{bmatrix}$$

<sup>(4)</sup> The Doolittle technique can also be used with  $X$  values.

The  $\mathbf{x}'\mathbf{x}$  matrix will be

$$\mathbf{x}'\mathbf{x} = \begin{bmatrix} \sum x_2^2 & \sum x_2x_3 & \sum x_1x_2 \\ \sum x_2x_3 & \sum x_3^2 & \sum x_1x_3 \\ \sum x_1x_2 & \sum x_1x_3 & \sum x_1^2 \end{bmatrix}$$

and, symbolically, its inverse is

$$(\mathbf{x}'\mathbf{x})^{-1} = \begin{bmatrix} d_{22} & d_{32} & d_{12} \\ d_{23} & d_{33} & d_{13} \\ d_{21} & d_{31} & d_{11} \end{bmatrix}$$

The Doolittle solution method is a convenient one because most of the statistics associated with the multiple regression problem may be written down almost directly once  $\mathbf{x}'\mathbf{x}$  and  $(\mathbf{x}'\mathbf{x})^{-1}$  are known. To illustrate,

$$b_{12.3} = \frac{-d_{12}}{d_{11}}$$

$$b_{13.2} = \frac{-d_{13}}{d_{11}}$$

$$s_{b_{12.3}}^2 = \frac{d_{11}d_{22} - d_{12}d_{21}}{(n-m)d_{11}^2}$$

$$s_{b_{13.2}}^2 = \frac{d_{11}d_{33} - d_{13}d_{31}}{(n-m)d_{11}^2}$$

$$r_{12.3}^2 = b_{12.3}^2 \left( \frac{d_{11}}{d_{22}} \right)$$

$$r_{13.2}^2 = b_{13.2}^2 \left( \frac{d_{11}}{d_{33}} \right)$$

Also

$$r_{1(23)}^2 = 1 - \frac{1}{(\sum x_1^2)d_{11}}$$

and

$$s_{1.23}^2 = \frac{1}{(n-m)d_{11}}$$

since

$$\sum x_1^2 = \frac{1}{d_{11}}$$

A full presentation of the Doolittle method will also give a convenient way of inverting  $\mathbf{x}'\mathbf{x}$  that requires a minimum of pencil work and provides a check after each of the steps in the inversion procedure. We shall not pursue this matter here.<sup>(5)</sup>

## AI6.4 THE GAUSS-MARKOV THEOREM

A very famous result in mathematical statistics that offers justification of the method of least squares is the Gauss-Markov theorem. Without

<sup>(5)</sup> See Dudley J. Cowden, "Correlation Concepts and the Doolittle Solution," *Journal of the American Statistical Association*, Vol. 38 (September, 1943), pp. 327-334.



proof, we state the theorem as follows:<sup>(6)</sup>

The method of least squares will offer an estimate of  $\beta$  in the model  $Y = X\beta + \epsilon$  which is the minimum variance *linear* unbiased estimate if  $E(\epsilon) = 0$  and  $E(\epsilon\epsilon') = \sigma^2 I$ .

The theorem is quite powerful in that it assures us, without the requirement that we state the exact distribution of the error term, that the least squares estimation technique will give a linear estimate of  $\beta$  that is unbiased and is "best," in the minimum variance sense.<sup>(7)</sup> It can also be shown that these properties hold for linear combinations of the estimate and  $\beta$ . Of course, we assume that  $X'X$  is nonsingular.

---

<sup>(6)</sup> A proof can be found in Franklin A. Graybill, *An Introduction to Linear Statistical Models* (New York: McGraw-Hill Book Company, Inc., 1961), pp. 115-117.

<sup>(7)</sup> The phrase "best linear unbiased estimator" is often abbreviated by the word BLUE.

## Tests of Homogeneity and Independence

Data are said to be homogeneous if all observations or all samples are governed by the same cause system. Different samples may be homogeneous with respect to their variances, but heterogeneous with respect to their means, and vice versa. One can test specific items, means, standard deviations, percentages, etc. to see whether they differ too much from the others to be attributed to chance. These tests are said to be specific. Alternatively, one can test the data as a whole. These tests are said to be general.

Data are said to be independent if the distributions of items in the different categories, say a contingency table, are unrelated. In correlation, if  $X_1$  and  $X_2$  are independent, then  $\rho = 0$ . But correlation, though it is a test of independence, is not a test of homogeneity. For a contingency table, however, if we have independence, then the data are homogeneous.

In this chapter we devote our attention to some general tests of homogeneity and independence. In the first three sections we will concentrate on the use of the analysis of variance (ANOVA) and the chi square distribution as they relate to the general testing of homogeneity. In the remaining section we will devote our attention to contingency tables and the general testing of both homogeneity and independence.

### 17.1 TESTING HOMOGENEITY USING ONE-WAY ANOVA

Table 17.1 gives the warp-breaking strength in pounds of six samples of four items each of a cloth product. The samples are arranged in columns with one column for each sample.

**TABLE 17.1: WARP-BREAKING STRENGTH IN POUNDS OF SIX SAMPLES OF FOUR ITEMS EACH**

<i>Item number \ Sample number j</i>	(1)	(2)	(3)	(4)	(5)	(6)	<i>Row total</i>
(1)	70	68	66	67	71	62	
(2)	68	66	64	66	68	59	
(3)	68	66	63	65	66	59	
(4)	62	63	60	60	57	56	
Column total $\bar{X}_j$	268 67.00	263 65.75	253 63.25	258 64.50	262 65.50	236 59.00	1540 385.00
Grand Mean							64.17

*Source: Dudley J. Cowden and William J. Connor, "The Use of Statistical Methods for Economic Control of Quality in Industry," Southern Economic Journal, Vol. 12 (Oct. 1945), pp. 115-129.*

To digress for a moment on notation, let us denote the individual items in Table 17.1 as  $X_{ij}$ . The subscripts on an  $X$  item refer to the row number and column number of the item, in that order. For example, in Table 17.1,  $X_{23} = 64$  but  $X_{32} = 66$ . We will also denote the total number of columns by  $k$  and the total number of rows by  $r$ . In Table 17.1,  $k = 6$  and  $r = 4$ . The total number of observations in the entire table will be denoted by  $N$ , and the total number of observations in any column will be denoted as  $n_j$ . To simplify the initial discussion, let us assume that all columns have the same number of observations. We will drop this assumption at the end of this section. In Table 17.1,  $N = 24$  and  $n = 4$  for all columns. Finally, a sample mean for a given column will be denoted  $\bar{X}_j$ , and the grand mean will be denoted  $\bar{X}$ . In Table 17.1,  $\bar{X}_1 = 67.00$  and  $\bar{X} = 64.17$ .

Each of the  $\bar{X}_j$  sample means in Table 17.1 is an estimate of a population mean  $\mu_j$ . The grand sample mean  $\bar{X}$  is an estimate of a grand population mean  $\mu$ . We wish to test

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_6$$

$$H_1: \mu_1, \mu_2, \dots, \text{ and } \mu_6 \text{ are not all equal}$$

The method of analysis follows the technique of partitioning variation analogous to that already encountered in regression analysis. The test statistic will be the  $F$  ratio.

$$F = \frac{s_1^2}{s_2^2} \quad (17-1)$$

where  $s_1^2$  and  $s_2^2$  are two independent estimates of the population variance.

**Model.** The population mean for each column is the grand population mean plus a column effect  $K_j$ . The relationship may be stated symbolically as

$$\mu_j = \mu + K_j \quad (17-2)$$

Thus, if  $\mu = 50$  and  $K_1 = -2$ ,  $K_2 = 5$ , and  $K_3 = -3$ , then  $K_1 + K_2 + K_3 = 0$ , and the column means are  $\mu_1 = 48$ ,  $\mu_2 = 55$ , and  $\mu_3 = 47$ . An  $X_{ij}$  value for any column is the column mean plus a random element  $E_{ij}$ .

$$X_{ij} = \mu_j + E_{ij}$$

Substituting  $X_{ij} - E_{ij}$  for  $\mu_j$  in Eq. (17-2), we obtain

$$X_{ij} = \mu + K_j + E_{ij} \quad (17-3)$$

Thus each  $X_{ij}$  value is assumed to be made up of three additive components: the grand population mean  $\mu$ , the column effect  $K_j$ , and a random element  $E_{ij}$ . We assume that the  $X_{ij}$  values in any column are normally distributed with means  $\mu_j$  and common variance  $\sigma^2$ . Thus, we assume that the population variances are homogeneous.<sup>(1)</sup>

$$\sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2 = \sigma^2$$

It is often possible to put the data in approximately normal form by using the logarithms of the data or by using some other transformation. However, the  $F$  test is a "robust" test, and mild departures from normality have only a minor effect.

Equation (17-3) can also be stated in deviation form:

$$(X_{ij} - \mu) = (\mu_j - \mu) + (X_{ij} - \mu_j) \quad (17-4)$$

**Symbolic Statement of the Test.** As we have already stated, we wish to test the null hypothesis that all population column means are equal against the alternative that they are not all equal. Another way of stating the null hypothesis is to say that the column effect for every column is zero. The alternative hypothesis is that not all column effects are zero. Thus

$$H_0: K_1 = K_2 = \dots = K_k = 0$$

$$H_1: K_1, K_2, \dots, \text{ and } K_k \text{ are not all zero}$$

If we substitute in Eq. (17-4) the estimates of  $\mu$  and  $\mu_j$  that are obtained from the sample, we have

$$(X_{ij} - \bar{X}) = (\bar{X}_j - \bar{X}) + (X_{ij} - \bar{X}_j)$$

---

<sup>(1)</sup> An alternative way of stating these assumptions is to specify the distribution of  $E_{ij}$ , where  $E_{ij}$  are independent random variables that are normally distributed with mean zero and variance  $\sigma^2$ .

If we square these deviations and sum them so that there are  $N$  squared deviations for each item, we have

$$\sum_{j=1}^k \sum_{i=1}^n (X_{ij} - \bar{X})^2 = n \sum_{j=1}^k (\bar{X}_j - \bar{X})^2 + \sum_{j=1}^k \sum_{i=1}^n (X_{ij} - \bar{X}_j)^2 \quad (17-5)$$

since  $N = nk$ . Again, we are assuming that all columns have the same number of observations  $n$ .

These measures of variation, or sums of squared deviations, or simply sums of squares, may be abbreviated as

$$\Sigma x_T^2 = \Sigma x_K^2 + \Sigma x_E^2 \quad (17-6)$$

and referred to verbally in this manner:

Total variation = column variation + error variation

The words "sums of squares" are often used instead of the word "variation," and "error variation" is sometimes called "within columns" variation.

Each of the measures of variation in Eq. (17-5) has an associated number of degrees of freedom that may also be put in equation form.

$$(N - 1) = (k - 1) + (N - k) \quad (17-7)$$

It is easy to see why these numbers of degrees of freedom are associated with the elements in Eq. (17-5). For total variation,  $X_{ij}$  occurs  $N$  times and  $\bar{X}$  occurs once, so the number of degrees of freedom is  $N - 1$ . For column variation,  $\bar{X}_j$  occurs  $k$  times and  $\bar{X}$  occurs once, so the number of degrees of freedom is  $k - 1$ . For error variation,  $X_{ij}$  occurs  $N$  times and  $\bar{X}_j$  occurs  $k$  times, and therefore the number of degrees of freedom is  $N - k$ . This fact can be put another way. There are  $n - 1$  degrees of freedom for each column, and there are  $k$  columns, so we have  $k(n - 1) = kn - k = N - k$  degrees of freedom altogether for error variation. The general rule is that an observation or statistic is limited in its freedom to vary by each restriction that is placed upon it. Thus we subtract from the number of values that are varying the number of restrictions imposed by the sample in order to determine the number of degrees of freedom to vary.

When any measure of variation is divided by its degrees of freedom, we obtain a measure of variance that is an unbiased estimate of the population variance  $\sigma^2$  if the null hypothesis is true. These measures of variance are often called *mean squares* and are

$$s_T^2 = \frac{\Sigma x_T^2}{N - 1}; \quad s_K^2 = \frac{\Sigma x_K^2}{k - 1}; \quad s_E^2 = \frac{\Sigma x_E^2}{N - k}$$

Of the three estimates,  $s_K^2$  and  $s_E^2$  are independent. It is possible to change the  $X_{ij}$  values in such a way that  $s_K^2$  is changed without changing  $s_E^2$ , and it is possible to change the  $X_{ij}$  values in such a way that  $s_E^2$  is changed without changing  $s_K^2$ , but in either case  $s_T^2$  will be affected.

TABLE 17.2: SYMBOLIC ANOVA TABLE

Source of variation	Amount of variation*	Degrees of freedom	Estimate of variance†
Total	$\sum x_T^2$	$N - 1$	
Columns	$\sum x_K^2$	$k - 1$	$s_K^2 = \sum x_K^2 / (k - 1)$
Error (within columns)	$\sum x_E^2$	$N - k$	$s_E^2 = \sum x_E^2 / (N - k)$

\* Also referred to as sums of squares, (S.S.).

† Also referred to as mean square, (M.S.). These are estimates of variance if the null hypothesis is true.

In the analysis of variance it is usually desirable to draw up an ANOVA table. This table is done symbolically in Table 17.2. Such a table enables us to observe the way in which the total variation is apportioned among its components. As the ruling of the table indicates, variation and degrees of freedom are additive, but variance is not.

To test the null hypothesis stated earlier, we compute

$$F_K = \frac{s_K^2}{s_E^2}$$

and evaluate the result by reference to Appendix 5.

**Computational Methods and Numerical Illustration.** Efficient formulas for computation of the three measures of variation are as follows:

$$\left. \begin{aligned} \sum x_T^2 &= \sum_{j=1}^k \sum_{i=1}^n X_{ij}^2 - C \\ \sum x_K^2 &= \frac{\sum_{j=1}^k \left( \sum_{i=1}^n X_{ij} \right)^2}{n} - C \\ \sum x_E^2 &= \sum x_T^2 - \sum x_K^2 \end{aligned} \right\} \quad (17-8)$$

and

$$C = \frac{\left( \sum_{j=1}^k \sum_{i=1}^n X_{ij} \right)^2}{N}$$

is usually referred to as the general correction term.

It is worthwhile to notice that

$$\sum x_E^2 = \sum_{j=1}^k \left[ \sum_{i=1}^n X_{ij}^2 - \frac{\left( \sum_{i=1}^n X_{ij} \right)^2}{n} \right]$$

For convenience, the data of Table 17.1 are repeated as Table 17.3, and the sums and squared sums are shown in the last two rows. Noting that

$$\sum_{j=1}^k \sum_{i=1}^n X_{ij}^2 = (70)^2 + (68)^2 + \cdots + (59)^2 + (56)^2 = 99,200.00$$

and using the values recorded in Table 17.3, we compute

$$\frac{\sum_{j=1}^k \left( \sum_{i=1}^n X_{ij} \right)^2}{n} = \frac{395,906}{4} = 98,976.50$$

and

$$C = \frac{\left( \sum_{j=1}^k \sum_{i=1}^n X_{ij} \right)^2}{N} = \frac{(1540)^2}{24} = 98,816.67$$

**TABLE 17.3: COMPUTATIONS FOR ANALYSIS OF VARIANCE (DATA OF TABLE 17.1)**

<i>Sample number</i> <i>Item number</i>	(1)	(2)	(3)	(4)	(5)	(6)	<i>Row total</i>
(1)	70	68	66	67	71	62	
(2)	68	66	64	66	68	59	
(3)	68	66	63	65	66	59	
(4)	62	63	60	60	57	56	
Column total (Column total) <sup>2</sup>	268 71,824	263 69,169	253 64,009	258 66,564	262 68,644	236 55,696	1,540 395,906

The measures of variation are then easily obtained.

$$\Sigma x_T^2 = 99,200.00 - 98,816.67 = 383.33$$

$$\Sigma x_K^2 = 98,976.50 - 98,816.67 = 159.83$$

$$\Sigma x_E^2 = 383.33 - 159.83 = 223.50$$

Table 17.4 is the analysis of variance table. It follows the format of Table 17.2.

**TABLE 17.4: ANALYSIS OF VARIANCE TABLE (DATA OF TABLE 17.1)**

<i>Source of variation</i>	<i>Amount of variation</i>	<i>Degrees of freedom</i>	<i>Estimate of variance</i>
Total	383.33	23	...
Columns	159.83	5	31.97
Error	223.50	18	12.42

We now compute the variance ratio

$$F_K = \frac{31.97}{12.42} = 2.574$$

With degrees of freedom  $\nu_1 = 5$  and  $\nu_2 = 18$ , the value of  $F$  at the 0.05 level is 2.773. The hypothesis that  $\mu_1 = \mu_2 = \dots = \mu_k$  is therefore accepted if

$\alpha = 0.05$ . Notice again that the rejection region is all located in the upper tail of the  $F$  distribution, since the null hypothesis will be rejected if  $F_K$  is large.

**Unequal Sample Sizes.** This test can also be used when the number of observations in each column is not the same. We need only to substitute  $n_j$  for  $n$  in each formula, and in Eq. (17-8) use

$$\frac{\sum_{j=1}^k \left( \sum_{i=1}^{n_j} X_{ij} \right)^2}{n_j} \quad \text{instead of} \quad \frac{\sum_{j=1}^k \left( \sum_{i=1}^n X_{ij} \right)^2}{n}$$

**Special Case When There Are Two Columns.** In Sec. 11.5 we tested the significance of difference between two independent sample means by using a  $t$  statistic defined as

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}}$$

We found  $t = 17.2$  by using the data of that section. We can also use the method of analysis of variance in this problem. Using the summary data defined in Sec. 11.5, we find that

$$\begin{aligned} \sum_{j=1}^k \sum_{i=1}^{n_j} X_{ij} &= n_1 \bar{X}_1 + n_2 \bar{X}_2 = 6932 + 4383 = 11,315 \\ C &= \frac{\left( \sum_{j=1}^k \sum_{i=1}^{n_j} X_{ij} \right)^2}{N} = \frac{(11,315)^2}{100} = 1,280,292.25 \\ \sum_{j=1}^k \sum_{i=1}^{n_j} X_{ij}^2 &= \left[ \sum x_1^2 + \frac{(\sum X_1)^2}{n_1} \right] + \left[ \sum x_2^2 + \frac{(\sum X_2)^2}{n_2} \right] \\ &= 972,720 + 394,181 = 1,366,901 \\ \sum x_T^2 &= 1,366,901 - 1,280,292.25 = 86,608.75 \\ \sum x_E^2 &= \sum x_1^2 + \sum x_2^2 = 21,635 \\ \sum x_K^2 &= 86,608.75 - 21,635 = 64,974 \\ F_K &= \frac{64,974/1}{21,635/99} = 296 \end{aligned}$$

Now it is always true that  $t = \sqrt{F}$  when  $\nu_1 = 1$ . Therefore,

$$t = \sqrt{296} = 17.2$$

This is the same value of  $t$  that was obtained in Sec. 11.5. The two methods are equivalent. The method of Sec. 11.5 is somewhat easier, and published  $t$ -tables are more detailed than published  $F$  tables.



## 17.2 TESTING HOMOGENEITY BY USING TWO-WAY ANOVA

Five makes of cars were driven for 4 weeks, and a different brand of fuel was used each week. A record was kept of miles per gallon, and the results are shown in Table 17.5. We wish to know whether there is a significant difference among cars and also among fuels.

**TABLE 17.5: FUEL CONSUMPTION OF FIVE MAKES OF CARS USING FOUR BRANDS OF GASOLINE**

<i>Fuel</i>	CAR					<i>Row total</i>	<i>Row total squared</i>
	(1)	(2)	(3)	(4)	(5)		
(1)	16.1	17.0	10.7	15.5	11.5	70.8	5,012.64
(2)	17.0	16.9	10.3	15.6	12.0	71.8	5,155.24
(3)	16.7	17.4	10.5	14.8	11.8	71.2	5,069.44
(4)	17.8	19.1	11.7	15.4	11.8	75.8	5,745.64
Column total	67.6	70.4	43.2	61.3	47.1	289.6	20,982.96*
Column total squared, or summary statistic	4569.76	4956.16	1866.24	3757.69	2218.41	17,368.26*	83,868.16*

$$* 83,868.16 = (289.6)^2 = \left( \sum_{j=1}^k \sum_{i=1}^r X_{ij} \right)^2, 17,368.26 = \sum_{j=1}^k \left( \sum_{i=1}^r X_{ij} \right)^2 \text{ and}$$

$$20,982.96 = \sum_{i=1}^r \left( \sum_{j=1}^k X_{ij} \right)^2.$$

*Source: Freely adapted from data provided by Dail Frazier, The Standard Oil Company (Ohio).*

For a problem of this type, there is another complication to consider. If a given car is driven by the same driver for 4 weeks, and a different brand of gasoline is used each week, then differences among weeks must be considered. Also, it is difficult to say whether differences in mileage are attributable to differences among cars or differences among drivers. If 20 different cars are each driven by 20 different drivers, all at one time, the statistical procedure is simplified, though random variability is increased, as well as the cost of the experiment. The procedure used in the present illustration is appropriate for this latter type of experiment.

**Model.** Three types of models are distinguishable in analysis of variance: random, fixed, and mixed. In the present case, if we select five

makes of cars at random, and four brands of gasoline of interest we have a mixed model. The results of the analysis are applicable to cars in general, but only to the brands of gasoline selected. The type of model does not affect the computations for tests of significance, but it affects the interpretation of the results.

The population mean for each cell,  $\mu_{ij}$ , is the grand population mean  $\mu$  plus the row effect  $R_i$  plus the column effect  $K_j$ . The relationship may be stated symbolically as

$$\mu_{ij} = \mu + R_i + K_j \quad (17-9)$$

Thus, if  $\mu = 50$  and there are two rows with  $R_1 = 4$  and  $R_2 = -4$  and if there are three columns with  $K_1 = -2$ ,  $K_2 = 5$ , and  $K_3 = -3$ , the cell means are as in Table 17.6. In this table

$\mu_{i.}$  denotes mean of row  $i$  for all columns

$\mu_{.j}$  denotes mean of column  $j$  for all rows

$\mu$  denotes mean of all  $X_{ij}$  values

**TABLE 17.6: ARITHMETIC ILLUSTRATION OF ANALYSIS OF VARIANCE MODEL: TWO BASES OF CLASSIFICATION**

$R_i \backslash K_j$	$K_1 = -2$	$K_2 = 5$	$K_3 = -3$	Row mean
$R_1 = 4$	$\mu_{11} = 52$	$\mu_{12} = 59$	$\mu_{13} = 51$	$\mu_{1.} = 54$
$R_2 = -4$	$\mu_{21} = 44$	$\mu_{22} = 51$	$\mu_{23} = 43$	$\mu_{2.} = 46$
Column mean	$\mu_{.1} = 48$	$\mu_{.2} = 55$	$\mu_{.3} = 47$	$\mu = 50$

It is assumed that there is no interaction between columns and rows; the column effect is constant from row to row, and the row effect is constant from column to column. Notice in Table 17.6 that there is a constant difference of 8 between the row entries; there is a constant difference of  $-7$  between the entries of column 1 and column 2, and a constant difference of 8 between column 2 and column 3.

An  $X_{ij}$  value for any cell is the cell mean plus a random element.

$$X_{ij} = \mu_{ij} + E_{ij}$$

Substituting  $X_{ij} - E_{ij}$  for  $\mu_{ij}$  in Eq. (17-9), we obtain

$$X_{ij} = \mu + R_i + K_j + E_{ij}$$

or

$$(X_{ij} - \mu) = R_i + K_j + (X_{ij} - \mu_{ij}) \quad (17-10)$$

From Table 17.6 it is apparent that  $R_i = \mu_{i.} - \mu$ ,  $K_j = \mu_{.j} - \mu$  and

$\mu_{ij} = \mu_{i.} + \mu_{.j} - \mu$ . Therefore, Eq. (17-10) can also be written in deviation form.

$$(X_{ij} - \mu) = (\mu_{i.} - \mu) + (\mu_{.j} - \mu) + (X_{ij} - \mu_{i.} - \mu_{.j} + \mu) \quad (17-11)$$

As before, we have the assumption of normality and uniformity of variance.

**Symbolic Statement of Test.** The null hypothesis is that the row effect is zero, *and* that the column effect is zero.

$$R_1 = R_2 = \dots = R_r = 0; \quad K_1 = K_2 = \dots = K_k = 0$$

An equivalent statement is

$$\begin{aligned} \mu_{1.} &= \mu_{2.} = \dots = \mu_{r.} = \mu \\ \mu_{.1} &= \mu_{.2} = \dots = \mu_{.k} = \mu \end{aligned}$$

If the sample estimates of  $\mu$ ,  $\mu_{i.}$ , and  $\mu_{.j}$  are substituted in Eq. (17-11), we obtain

$$(X_{ij} - \bar{X}) = (\bar{X}_{i.} - \bar{X}) + (\bar{X}_{.j} - \bar{X}) + (X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X})$$

Again, if equal numbers of observations in the columns are assumed, the components of variation are the sums of squares of these quantities.

$$\begin{aligned} \sum_{j=1}^k \sum_{i=1}^r (X_{ij} - \bar{X})^2 &= k \sum_{i=1}^r (\bar{X}_{i.} - \bar{X})^2 + r \sum_{j=1}^k (\bar{X}_{.j} - \bar{X})^2 \\ &\quad + \sum_{j=1}^k \sum_{i=1}^r (X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X})^2 \quad (17-12) \end{aligned}$$

and may be abbreviated as

$$\Sigma x_T^2 = \Sigma x_R^2 + \Sigma x_K^2 + \Sigma x_E^2$$

and stated verbally as

Total variation = row variation + column variation + error variation

The degrees of freedom for the different measures of variation given in Eq. (17-12) are additive.

$$(N - 1) = (r - 1) + (k - 1) + (N - r - k + 1)$$

$$\text{or} \quad (N - 1) = (r - 1) + (k - 1) + (r - 1)(k - 1) \quad (17-13)$$

since  $N = rk$  when there are equal numbers of observations in each column.

The estimates of variance are

$$s_R^2 = \frac{\Sigma x_R^2}{r - 1} \quad s_K^2 = \frac{\Sigma x_K^2}{k - 1} \quad s_E^2 = \frac{\Sigma x_E^2}{(r - 1)(k - 1)}$$

and the variance ratios are

$$F_R = \frac{s_R^2}{s_E^2} \quad F_K = \frac{s_K^2}{s_E^2}$$

**Computational and Numerical Illustration.** For efficient computation, the formulas are

$$\left. \begin{aligned} \Sigma x_T^2 &= \left( \sum_{j=1}^k \sum_{i=1}^r X_{ij}^2 \right) - C \\ \Sigma x_R^2 &= \frac{\sum_{i=1}^r \left( \sum_{j=1}^k X_{ij} \right)^2}{k} - C \\ \Sigma x_K^2 &= \frac{\sum_{j=1}^k \left( \sum_{i=1}^r X_{ij} \right)^2}{r} - C \\ \Sigma x_E^2 &= \Sigma x_T^2 - \Sigma x_R^2 - \Sigma x_K^2 \end{aligned} \right\} \quad (17-14)$$

and

$$C = \frac{\left( \sum_{j=1}^k \sum_{i=1}^r X_{ij} \right)^2}{N}$$

Again, as in the last section,  $C$  is referred to as the general correction term. From Table 17.5 we compute

$$\begin{aligned} C &= \frac{\left( \sum_{j=1}^k \sum_{i=1}^r X_{ij} \right)^2}{N} = \frac{83,868.16}{20} = 4193.41 \\ \sum_{j=1}^k \sum_{i=1}^r X_{ij}^2 &= (16.1)^2 + (17.0)^2 + \cdots + (11.8)^2 + (11.8)^2 \\ &= 4348.38 \\ \Sigma x_T^2 &= 4348.38 - 4193.41 = 154.97 \\ \Sigma x_R^2 &= \frac{20,982.96}{5} - 4193.41 = 3.18 \\ \Sigma x_K^2 &= \frac{17,368.26}{4} - 4193.41 = 148.66 \\ \Sigma x_E^2 &= 154.97 - 3.18 - 148.66 = 3.13 \end{aligned}$$

Table 17.7 is the ANOVA table. The value of the computed  $F$  ratios are given below.

**TABLE 17.7: ANALYSIS OF VARIANCE TABLE (DATA OF TABLE 17.5)**

Source of variation	Amount of variation	Degrees of freedom	Estimate of variance
Total	154.97	19	...
Rows (fuels)	3.18	3	1.060
Columns (cars)	148.66	4	37.165
Error	3.13	12	0.2608

$$F_R = \frac{1.060}{0.2608} = 4.064; \quad F_K = \frac{37.165}{0.2608} = 142.5$$

Entering Appendix 5, we find, with  $\alpha = 0.05$ ,  $\nu_1 = 3$ ,  $\nu_2 = 12$ , that the upper rejection value of  $F$  is 3.490. Since  $F_R = 4.064 > 3.49$ , we reject the null hypothesis that all row effects are zero. Similarly, with  $\alpha = 0.05$ ,  $\nu_1 = 4$ ,  $\nu_2 = 12$ , the upper rejection value for  $F$  is 4.26. We reject the null hypothesis that all column effects are zero, since  $F_K = 142.5 > 4.26$ . Thus, at the 0.05 level, there is a significant difference among cars and among fuels. The difference among cars is especially convincing.

**Special Case Where There Are Two Columns.** In Sec. 11.5 we tested the significance of mean difference between 10 pairs of test cubes of concrete. We formed a  $t$  statistic defined by

$$t = \frac{\bar{D}}{s/\sqrt{n}}$$

and found  $t = 2.76$ . It is also possible to use analysis of variance. Using the data of Table 11.1, we obtain

$$\sum x_T^2 = 2846.8$$

$$\sum x_R^2 = 1842.8$$

$$\sum x_K^2 = 460.8$$

$$\sum x_E^2 = 543.2$$

Then

$$F_K = \frac{460.8/1}{543.2/9} = 7.63$$

which is the square of the  $t$  statistic computed in Sec. 11.5.

## 17.3 TESTING HOMOGENEITY USING CHI SQUARE: GOODNESS OF FIT

If data are homogeneous they should conform reasonably well to some probability distribution. The chi square test offers a method of examining how well a theoretically generated frequency distribution describes, or fits, an observed frequency distribution.

The following is a simple illustration of the use of the  $\chi^2$  distribution for this purpose. After being blindfolded, each of 60 persons was given three cigarettes, each of a different brand not disclosed to him, and asked to state which he liked best. If there is no difference among the brands, each brand should receive 20 first-choice votes, i.e., one-third of the votes. The actual votes were as given in Table 17.8, column  $f_i$ . The theoretical number of votes, on the assumption of no difference among the brands, is given in column  $\hat{f}_i$ .

**TABLE 17.8: THEORETICAL AND OBSERVED PREFERENCE FOR THREE BRANDS OF CIGARETTES, AND COMPUTATION OF THE APPROXIMATION TO  $\chi^2$** 

<i>Brand</i>	$f_i$	$\hat{f}_i$	$f_i - \hat{f}_i$	$(f_i - \hat{f}_i)^2$	$(f_i - \hat{f}_i)^2 / \hat{f}_i$
A	24	20	4	16	0.8
B	17	20	-3	9	0.45
C	19	20	-1	1	0.05
Total	60	60	0	...	$1.30 = X^2$

The null hypothesis is that the distribution of  $f$  and  $\hat{f}$  are the same. The alternative hypothesis is that the distributions are not the same. Under the null hypothesis the distribution of the statistic

$$X^2 = \sum_{i=1}^r \frac{(f_i - \hat{f}_i)^2}{\hat{f}_i} \quad (17-15)$$

approaches the  $\chi^2$  distribution as the sample size increases.<sup>(2)</sup> The smallest value that the statistic given in Eq. (17-15) could take is zero, in which case the observed and theoretical frequencies would be identical. On the other hand, if all votes had been cast for a single brand, we would have obtained the greatest possible discrepancy between the observed and theoretical frequencies, and the calculated value of the statistic would have been 840. Obviously, the greater the positive departure of the statistic from zero, within the possible range, the greater the discrepancy between the theoretical and observed frequencies.

In Table 17.8 we notice that the observed and theoretical frequencies have been made to have the same sum. Thus, one restriction has been placed on the three classes containing the three theoretical frequencies, and only two classes are free to vary. Using Appendix 6,  $r - 1 = 2$  degrees of freedom, and  $\alpha = 0.05$ , we find the rejection value of  $\chi^2$  to be about 5.991. Notice again that all of the rejection region is placed in the upper tail of the  $\chi^2$  distribution, since the null hypothesis is to be rejected only if  $X^2$  is large. The calculated value of  $X^2$ , given in Table 17.8, is 1.30 and does not exceed the rejection value of  $\chi^2$  at the 0.05 level. The hypothesis of homogeneity must, therefore, be accepted.

The chi square distribution may be used to examine the goodness of fit of other distributions as well. If the hypothesis was that a frequency distribution was normally distributed, the chi square distribution would be used. The number of degrees of freedom would be the number of classes minus three, since the normal frequencies were made to conform to the observed distribution in three respects: number of frequencies, mean, and standard deviation (see the appendix to Chapter 7).

<sup>(2)</sup> The exact test using the multinomial distribution is beyond the scope of this text.

**Special Cases When There Are Two Classes.** In Sec. 12.2, another taste test was illustrated. Each of 44 persons was given a drink of R-C Cola® and a drink of Coca-Cola® and asked to state which was which. Of the 44 persons, 34 answered correctly. In Sec. 12.2 the test statistic applied was

$$z = \frac{d - nP}{\sigma_d}$$

and the value of  $z$  was found to be 3.62.

This problem can also be handled by use of chi square, as illustrated in Table 17.9. Here, the value of  $X^2$  is found to be 13.090 and its square root is

**TABLE 17.9: COMPUTATION OF APPROXIMATION TO  $\chi^2$  FOR TEST OF HOMOGENEITY: TWO CLASSES**

<i>Answer</i>	$f_i$	$\hat{f}_i$	$f_i - \hat{f}_i$	$(f_i - \hat{f}_i)^2$	$(f_i - \hat{f}_i)^2 / \hat{f}_i$
Correct	34	22	12	144	6.545
Incorrect	10	22	-12	144	6.545
Total	44	44	0	...	$X^2 = 13.090$

3.62, which is the value of  $z$  found previously. Comparison of Appendix 3 and Appendix 6 will aid in understanding why this is true. Notice that when  $\nu = 1$

$$(z_{Q/2})^2 = \chi_{Q,1}^2$$

For example, at  $\alpha = 0.05$ ,  $z_{0.025} = 1.95996$  and  $(1.95996)^2 = 3.841 = \chi_{0.05,1}^2$  when  $\nu = 1$ . Hence, the  $\chi^2$  test and the two-sided  $z$  test are equivalent when  $\nu = 1$ .

**Correction for Continuity.** Table 17.8 represents a  $3 \times 1$  ("3 by 1") table, since there are three rows and one column for the expected frequencies. Table 17.9 is a  $2 \times 1$  table. Whenever the number of degrees of freedom for the  $X^2$  statistic is one, and this is the case in our  $2 \times 1$  table, a correction for continuity, known as Yates' correction, is sometimes applied to  $X^2$  in hopes of improving the approximation to the  $\chi^2$  distribution. The correction amounts to subtracting 0.5 from each absolute value of  $f_i - \hat{f}_i$  before squaring.

$$\tilde{X}^2 = \sum_{i=1}^r \frac{(|f_i - \hat{f}_i| - 0.5)^2}{\hat{f}_i} \quad (17-16)$$

The student can verify that the corrected value of  $X^2$  as calculated from Table 17.9 is 12.02, which, again, is the square of the  $z$  statistic when the  $z$

statistic is corrected for continuity as illustrated in Sec. 12.2, note 3. However, as we noted in Sec. 12.2, the correction often overcorrects.

### 17.4 TESTING HOMOGENEITY AND INDEPENDENCE USING CHI SQUARE: CONTINGENCY TABLES

Twenty salesmen have been rated by their superior as having good or poor selling ability and, as well, have been given a psychological test to measure their capacity to modify their routine activities (the Downey Will-Temperament Test for Flexibility). The results are given in Table 17.10. This table is called a  $2 \times 2$  contingency table.

TABLE 17.10: FLEXIBILITY AND SELLING ABILITY OF TWENTY SALESMEN

<i>Selling ability</i>	FLEXIBILITY		<i>Total</i>
	<i>Poor</i>	<i>Good</i>	
Good	3	7	10
Poor	8	2	10
Total	11	9	20

Source: Kornagy and Graham, *The Selection and Training of Salesmen* (New York: McGraw-Hill Book Co., Inc. 1925), p. 294.

It is apparent at a glance that the better salesmen are superior in flexibility. Only 5 out of 20 salesmen are exceptions to this general tendency. In symbolic form our observed frequencies are

<i>Selling ability</i>	FLEXIBILITY		<i>Total</i>
	<i>Poor</i>	<i>Good</i>	
Good	$f_{11}$	$f_{12}$	$n_{1.}$
Poor	$f_{21}$	$f_{22}$	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$N$

The null hypothesis is that the frequencies for the cells in a given row are independent of the column in which they are located, and that the frequencies in a given column are independent of the row in which they are located. Under this hypothesis the expected frequencies for any cell may be found by use of

$$f_{ij} = \frac{(n_{i.})(n_{.j})}{N} \quad (17-17)$$



Then, for our illustration, the expected frequencies, under the assumption of independence, are:

$\hat{f}_{11} = \frac{10(11)}{20} = 5.5$	$\hat{f}_{12} = \frac{10(9)}{20} = 4.5$	$n_{1.} = 10$
$\hat{f}_{21} = \frac{10(11)}{20} = 5.5$	$\hat{f}_{22} = \frac{10(9)}{20} = 4.5$	$n_{2.} = 10$
$n_{.1} = 11$	$n_{.2} = 9$	$N = 20$

The test statistic is computed in the usual manner.

$$X^2 = \sum_{j=1}^k \sum_{i=1}^r \frac{(f_{ij} - \hat{f}_{ij})^2}{\hat{f}_{ij}} \quad (17-18)$$

which, for our data, is

$$\begin{aligned} X^2 &= \frac{(3 - 5.5)^2}{5.5} + \frac{(8 - 5.5)^2}{5.5} + \cdots + \frac{(2 - 4.5)^2}{4.5} \\ &= 5.050 \end{aligned}$$

Since both the row and column sums of the theoretical frequencies have been made to agree with the row and column sums of the observed frequencies, the number of degrees of freedom for a contingency table is

$$(N - 1) - (r - 1) - (k - 1)$$

which reduces to

$$(r - 1)(k - 1) \quad (17-19)$$

since  $N = rk$ . In the present case we have a  $2 \times 2$  table, so there is a single degree of freedom. We may enter either the normal probability table using  $\sqrt{X^2}$  or the chi-square table with one degree of freedom. The student may verify that in either case the null hypothesis of independence will be rejected at the 0.05 level. The alternative hypothesis that selling ability depends partly on flexibility is accepted.

**Correction for Continuity.** For a  $2 \times 2$  table it is sometimes advised that the statistic  $X^2$  be corrected for continuity. James E. Grizzle has said, however, that the correction for continuity is not desirable, in that one is more likely to control the probability of making a type I error (at least at the 0.05 level) when the uncorrected statistic is used.<sup>(3)</sup> Exact tests based

<sup>(3)</sup> James E. Grizzle, "Continuity Correction in the  $\chi^2$ -test for  $2 \times 2$  Tables," *The American Statistician*, October, 1967, pp. 28-32. See also: Nathan Mantel and Samuel W. Greenhouse, "What is the Continuity Correction?," *The American Statistician*, December, 1968, pp. 27-30.

upon the hypergeometric distribution are possible but are more difficult.<sup>(4)</sup> Complete reliance should not be placed upon the  $X^2$  statistic when the smallest expected frequency is less than five.

**General Contingency Table.** The analysis of an  $r \times k$  contingency table is simply an extension of the  $2 \times 2$  table. The expected frequencies are calculated by using Eq. (17-17); the approximation to chi square is calculated by using Eq. (17-18), and the number of degrees of freedom is found by using Eq. (17-19). Again, warnings concerning small expected frequencies are in order.

## PROBLEMS

1. Gas-operated rifles have in the barrel a small gas port that allows part of the gas associated with the passage of a given bullet to feed back into the rifle and operate the action of the weapon. The question is often raised as to whether or not the leakage of this gas affects the velocity of the bullet in any appreciable way. The following data, adapted from *The American Rifleman*, April, 1966, pertain to the M-14 rifle. (Assume that two batches of four rifles each were tested.)

Test number	VELOCITY (FPS) WITH GAS PORT	
	Open	Closed
1	2864	2860
2	2864	2876
3	2882	2872
4	2886	2880

Does the closing of the gas port cause a significant reduction in velocity? What are your assumptions in conducting this test? Use  $\alpha = 0.01$ .

2. A cattle feed lot operator is interested in the weight gain of yearling bulls of three different breeds. He is also interested in differences between four different kinds of prepared feeds. The feeds and cattle breeds are chosen because of interest, and thus the model is fixed. Four bulls are chosen at random from each breed and then randomly assigned to a prepared feed. All cattle are kept on the same feed lot and are fed for a three-week period. The weight gain figures in pounds

<sup>(4)</sup> E. S. Pearson and H. O. Hartley, eds., *Biometrika Tables for Statisticians*, (Cambridge: Cambridge University Press, 1966), pp. 71-78, discuss an exact test for  $2 \times 2$  tables.

are given below. Test the null hypothesis that the row effect is zero and the column effect is zero, using  $\alpha = 0.05$ . Formulate your conclusions in words.

<i>Feed \ Breed</i>	<i>Hereford</i>	<i>Angus</i>	<i>Charolais</i>
A	100	103	110
B	75	90	100
C	120	100	120
D	100	90	100

3. A magazine measured dollar retail sales per family and the circulation per 1000 families of its magazine for a sample 155 counties. The results are given below. Test for independence of circulation and retail sales at  $\alpha = 0.01$ . Formulate your conclusions in words.

<i>Circulation</i>	RETAIL SALES		<i>Total</i>
	<i>\$1000-3399</i>	<i>\$3400-5799</i>	
10.0-32.4	25	59	84
0.0-9.9	60	11	71
	85	70	155

4. Show that when there are two columns

$$(1) \quad t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}} = \sqrt{F_K}$$

$$(2) \quad t = \frac{D}{s_D} = \sqrt{F_K}$$

5. Flip a coin 50 times and record the frequency of heads and tails. Test the coin for "fairness," using chi square and  $\alpha = 0.01$ .

6. Pearson's contingency coefficient  $C$ . Show that the statistic

$$C = \sqrt{\frac{X^2}{X^2 + N}}$$

is zero when the rows and columns in a  $2 \times 2$  contingency table are independent. What is the maximum value that  $C$  may assume for a  $2 \times 2$  table?

7. Assume a  $2 \times 2$  table with the same number of observations in all rows and columns and a total number of observations  $N$ . Now, assume a second  $2 \times 2$  table with twice the number of observations in all rows and columns and a total number of observations  $2N$ . What is the relationship between the values of  $X^2$  for the two tables?

8. Show that for a  $2 \times 2$  table,  $X^2$  can be computed by using

$$X^2 = (f_{ij} - f_{ij})^2 \sum_{j=1}^k \sum_{i=1}^r \frac{1}{f_{ij}}$$

where  $(f_{ij} - f_{ij})^2$  is a constant (the same for all cells).

9. Tschuprow's  $T^2$ . If we define the mean square contingency in an  $r \times k$  table as

$$\phi^2 = \frac{X^2}{N}$$

then  $T^2$  is defined as

$$T^2 = \frac{\phi^2}{\sqrt{(r-1)(k-1)}}$$

Show that  $T^2$  cannot exceed one in a  $2 \times 2$  table. Show also that  $T^2$  must be less than one in a  $2 \times 3$  table.

## Index Numbers

Index numbers are devices for comparing the magnitude of groups of related variables. They may compare (1) changes occurring over time, (2) differences between places, and (3) differences between like categories, such as persons, organizations, or objects. An index number is obtained by combining the variables by means of a total or average. Usually each index number in a series is expressed as a percentage of some convenient base. When a series of index numbers is obtained, it is customary to refer to the series as an *index*.

The most common indexes measure variations over time. The total amount of money spent each year, relative to some base year, varies from year to year because of changes in the number of units of the different commodities bought, and because of changes in the unit prices of these commodities. Thus, we have three variables:

- $V$ , the value relative to some base year of the group of commodities, sometimes referred to as a value index number.
- $Q$ , the physical quantity relative to some base year of the commodities considered as a group, or the quantity index number.
- $P$ , the price relative to the base year of the commodities considered as a group, or the price index number.

The relationship between these three index numbers is, ideally,

$$V = PQ$$

This relationship seems reasonable when we consider that the amount of money spent for any commodity, say for cigarettes, is the price per package multiplied by the number of packages bought.

## 18.1 USES OF INDEX NUMBERS

1. Price movements are a guide to business policy. Rising prices call for one policy, whereas falling prices necessitate a different one. A stable or slowly rising price level is generally considered as a condition favorable to business stability. Stock market prices are considered especially significant, since changes in stock prices are sometimes forerunners of changes in other prices.

2. Indexes of industrial production are useful not only as indicators of business conditions, but also as a base with which to compare production in one's own business. For the latter purpose it is usual to compare production in the individual business with production in the industry.

3. Sometimes it is possible to combine series that are causally related to one's own business in such a way that changes in the index will forecast changes in one's own business.

4. It may be difficult to tell whether a company's sales are increasing in *physical volume* as opposed to *dollar value*. The products sold may be numerous and diverse; some may be increasing while other are decreasing in regard to the number of units sold. Likewise, the unit prices may be changing. An index of physical volume of company sales may therefore be constructed, or the dollar value of the sales may be "deflated," by dividing them by an index of company prices.

In Table 18.1, a series representing average hourly earnings received in

**TABLE 18.1 DEFLATION OF AVERAGE HOURLY EARNINGS IN U.S. MANUFACTURING INDUSTRIES, 1964-1967**

Year	Average earnings (dollars per hour)	Consumer Price Index (1957-59 = 100)	Estimated real average earnings per hour
1964	2.58	108.8	2.37
1965	2.66	111.0	2.40
1966	2.77	114.7	2.41
1967	2.91	118.2	2.46

*Source:* Federal Reserve Bulletin, various issues. Data are for December of the year given.

United States manufacturing industries is deflated by use of the consumer price index. The deflated series, real average hourly earnings, is obtained by dividing the nominal series by the price index. The deflated series is considered by most economists to be the more meaningful of the two, since it gives a better impression of the purchasing power of the wages and salaries received in terms of *real* goods and services.

5. An index of company prices may also provide favorable publicity in case company prices have fallen in comparison with company wage rates or other price indexes.

6. Many wage contracts contain an escalator clause providing for adjustment of wages on the basis of changes in the consumer price index, or CPI, as it is usually called. Usually these contracts call for adjustments quarterly. For example, if the hourly wage rate has been \$1.86 per hour and the CPI changes from 106.2 to 106.8, the new wage rate will be increased from \$1.86 to  $(106.8/106.2)1.86 = \$1.87$  per hour, if the escalator clause calls for adjustment in wage rates proportionate to changes in the CPI. Although a change in 1 cent per hour does not sound like much, the aggregate effect is considerable when it applies to millions of workers.

Sometimes other long-term contracts call for payments to be adjusted on the basis of the CPI. An example is the payment for expensive products manufactured to special order and taking a long time to make.

7. Index numbers are also used to adjust financial statements for the effect of price changes. Depreciation charges, especially, should be adjusted so that in time of rising prices enough money will be charged off to cover the cost of replacement of durable goods.

## 18.2 PROBLEMS IN INDEX NUMBER CONSTRUCTION

In this section we will discuss, in order, the selection of a base, the type of formula to use, the weighting system, and the selection of suitable data.

**Selection of Base.** Selection of a base with which to compare the various index numbers does not raise difficult theoretical questions. The choice depends on the purpose of the index. For a general index number a "normal" period should be chosen as base. Since comparisons among indexes are often made, one should also consider what base period has already been selected for related indexes. The United States Government has suggested that the three-year period 1957-59 is appropriate for many indexes, and there now seems to be a tendency to adopt this period as a base. Nevertheless, a number of different bases are in current use, and it does not seem that a given set of years is necessarily "normal" for all series of data or appropriate for all purposes.

**Type of Formula.** Selection of a type of formula that is technically sound and appropriate for the particular purpose in view is a subject which has occupied the attention of statisticians for decades and which has resulted in

divergent solutions. The problem is perhaps of greater theoretical than practical importance.

**Weighting System.** Weighting the different series according to their importance is a troublesome problem, especially if their importance is changing. The fact that the relative prices of different commodities partly determine the relative quantities consumed introduces a logical difficulty. It will be shown, however, that only approximate accuracy in weights is required.

**The Data for Index Numbers.** Too much emphasis cannot be put upon the practical problem of selecting the data that are the raw material of the index. Without doubt, the most important problem in index number construction is the selection of suitable data. The data must be accurate and comparable, and the sample must be adequate in size and as nearly representative as possible. The usefulness of an index is usually enhanced if it can be made available without delay. Hence, preference should be given to data taken from sources that report promptly. Naturally, the data to be used depend on what one is trying to measure. A wholesale price index requires wholesale prices; a consumer price index necessitates data not only on retail prices of food, clothing, and house furnishings, but on rents, gas and electric rates, and so on. The data should apply to those commodities used by the class of persons for whom the index is intended.

1. *Accuracy.* Although prices and quantities are likely to be accurate if gathered from the internal records of one's own company, one cannot always be sure of the accuracy of data reported by others. For example, the housewife may not keep accurate records of the quantities she purchases, and prices reported by shoppers at retail stores may be for commodities other than those specified for the price index. Furthermore, accurate price quotations are often impossible to obtain because of discounts, special sales, hidden charges, etc.

2. *Comparability.* Standard grades of the same commodity are, of course, comparable between different dates. But is a 1949 automobile comparable with a 1969 one? Generally, one would rather have a 1969 car than a 1949 model of the same general type. Likewise, any physical volume index that treats an old carbon filament clear glass lamp as the equivalent of a modern, nitrogen-filled, inside-frosted bulb contains a serious mistake.

There are at least two currently used methods of improving the comparability of price quotations of commodities. The first is called "linking in" and is used where possible. Assume sweater price quotations as follows:

<i>Specification</i>	<i>Year 0</i>	<i>Year 1</i>	<i>Year 2</i>
<i>A</i>	\$4.00	\$3.50	...
<i>B</i>	...	\$5.00	\$5.50



In stating the price of a sweater in year 2 relative to year 0, we would compute, not  $\$5.50/\$4.00 = 137.5$  percent, but

$$\frac{3.50}{4.00} \cdot \frac{5.50}{5.00} = 96.25 \text{ percent}$$

On the other hand, if sweater *A* were not available in year 1 or year 2, but it was known that sweater *B* cost \$1.50 more to manufacture than sweater *A*, then we could make a direct adjustment for cost.

$$\frac{5.00 - 1.50}{4.00} \cdot \frac{5.50}{5.00} = 96.25 \text{ percent}$$

The two methods of adjustment would not usually yield exactly the same results.

3. *Representativeness*. Since index numbers are usually obtained from samples, one must try to obtain a sample that behaves like the universe from which it is drawn. Probably the most satisfactory way to do this is to divide the original data into groups and subgroups and try to obtain proportionate representation in each group. Thus, if the value of farm products marketed is three times as large as that of processed foods, the value of the former items in the sample should be three times that of the latter for any index of prices or quantities. These large groups can each be split into smaller ones, such as fresh and dried fruits and vegetables, grains, etc., and the same procedure can be followed. In short, the problem is similar to that of stratified sampling. In practice, the weight attached to the commodities in any group is adjusted so that the value of the sample will be correct for each group; i.e., each commodity is weighted, not according to its own importance, but according to the importance of the group of commodities that it represents. Commodities selected to represent a group of commodities in a price index should, of course, be those which, taken together, are typical of the price movements of all the commodities in the group.

4. *Adequacy*. It has been pointed out that the reliability of the mean of a random sample increases with the square root of the *number* of items included. Likewise, the larger the *proportion* of items included, the more reliable is the mean. In index number construction reliability depends upon the proportion of total *value* included in the index. It would appear, then, that one should ordinarily select the most important items first and as many other suitable items as it seems worthwhile to include. The absolute number of items to use or the proportion of total value to include are questions that cannot be answered in general terms. The problem is further complicated because of these factors:

1. Weights are values, rather than frequencies.
2. The commodities are usually selected partly on the basis of judgment.
3. Weights are assigned to individual commodities or groups of commodities on a partly arbitrary basis.

4. The prices (and quantities) of the different commodities are interdependent.
5. Stratified sampling is used.

Thus is it almost impossible to obtain the standard error of an index number because of the use of judgment in selecting commodities and the arbitrary assignment of weights. The use of value weights and stratified sampling contribute to this problem, although they, in themselves, are not insuperable obstacles.

### 18.3 INDEX NUMBER SYMBOLS

Index numbers are computed from prices and/or quantities of individual commodities. These are designated by lower case letters.

$p$  is the price of an individual commodity.  
 $q$  is the quantity of an individual commodity.

Usually a subscript is attached to  $p$  or  $q$ . The subscript 0 is generally (though not always) used to refer to the base year, and 1, 2, etc., refer to other years in chronological order.

$p_0$  is the price of a commodity in the base year.  
 $p_1$  is the price of a commodity in year 1.  
 $p_2$  is the price of a commodity in year 2.  
 $p_j$  is the price of a commodity in year  $j$ .

Analogous meanings are attached to the symbols  $q_0$ ,  $q_1$ ,  $q_2$ , etc. Capital letters refer to index numbers.

$P$  means price index number.  
 $Q$  means quantity index number.

Subscripts attached to  $P$  or  $Q$  refer to the years being compared.

$P_{01}$  means price index for year 1 relative to year 0.  
 $P_{02}$  means price index for year 2 relative to year 0.  
 $P_{ij}$  means price index number for year  $j$  relative to year  $i$ .

Price and quantity index numbers are almost always expressed on a 100 percentage point basis. Such expression is implied in the formulas to follow.

### 18.4 SIMPLE INDEX NUMBERS

**Simple Aggregate Index Numbers.** The simplest, but one of the least satisfactory, methods of index number construction is to ascertain

**TABLE 18.2: RELATED VEGETABLE OIL PRICES, 1963-1966, AND SIMPLE AGGREGATIVE PRICE INDEX (PRICES ARE IN DOLLARS PER POUND.)**

Type of oil	1963	1964	1965	1966
	$p_0$	$p_1$	$p_2$	$p_3$
Soybean	0.13	0.12	0.13	0.14
Cottonseed	0.15	0.14	0.15	0.18
Linseed	0.13	0.13	0.13	0.13
Total	0.41	0.39	0.41	0.45
Index number*	100.0	95.1	100.0	109.8

\* Equation (18-1). Prices are taken to be exact.

Source: Office of Business Economics, Survey of Current Business, various issues.

the total cost in each year of buying one unit of each commodity to be included in the index and to express this total cost each year as a percentage of the base year cost. This operation is carried out for three vegetable oil products, whose prices are given in Table 18.2. The formula is

$$P_{01} = \frac{\sum p_1}{\sum p_0} \quad (18-1)$$

The simple aggregative index number assigns equal importance to the absolute change of each commodity. The result is that a commodity with a high price per unit tends to exert more influence on the simple aggregative index number than does a commodity with a low price per unit. In our present illustration cottonseed oil exerts a greater influence than soybean oil; yet, as can be seen from Table 18.3, over three times as much soybean oil was consumed in end products in 1963. The weighting is, in fact, haphazard and illogical: the unit price quotation for different commodities may be such units as gram, pound, ton, barrel, and so on. The fact that a particular unit of

**TABLE 18.3: UNITED STATES CONSUMPTION IN END PRODUCTS OF RELATED VEGETABLE OIL PRODUCTS, 1963-1966, AND SIMPLE AGGREGATIVE QUANTITY INDEX (QUANTITIES ARE IN MILLIONS OF POUNDS, MONTHLY AVERAGES.)**

Type of oil	1963	1964	1965	1966
	$q_0$	$q_1$	$q_2$	$q_3$
Soybean	322	368	367	433
Cottonseed	96	114	123	105
Linseed	32	31	19	19
Total	450	513	509	557
Index number*	100.0	114.0	113.1	123.8

\* Equation (18-2). Quantities are taken to be exact.

Source: Office of Business Economics, Survey of Current Business, various issues.

measure happens to be quoted commercially may have nothing to do with the economic importance of the good.

The analogous quantity index number formula is

$$Q_{0j} = \frac{\sum q_j}{\sum q_0} \quad (18-2)$$

This formula compares the cost in the given year with the cost in the base year of buying the goods actually bought in the given year if the price for each commodity in each year was \$1.00 per unit. This assumption is obviously unrealistic.

**Simple Average of Relatives Index Numbers.** Table 18.4 shows the prices of the different oils relative to 1963. These relative prices  $p_j/p_0$  are usually referred to as *price relatives*. A crude price index number can

**TABLE 18.4: RELATED VEGETABLE OIL PRICES RELATIVE TO 1963, AND SIMPLE ARITHMETIC AVERAGE OF RELATIVES INDEX (PERCENT)**

Type of oil	1963	1964	1965	1966
	$\frac{p_0}{p_0}$	$\frac{p_1}{p_0}$	$\frac{p_2}{p_0}$	$\frac{p_3}{p_0}$
Soybean	100.0	92.3	100.0	107.7
Cottonseed	100.0	93.3	100.0	120.0
Linseed	100.0	100.0	100.0	100.0
Total	300.0	285.6	300.0	327.7
Index number*	100.0	95.2	100.0	109.2

\* Equation (18-3).

Source: Table 18.2.

be obtained by simply averaging the price relatives for each year. This type of index number is known as the *simple average of price relatives* and is indicated by

$$P_{0j} = \frac{1}{n} \sum \left( \frac{p_j}{p_0} \right) \quad (18-3)$$

The chief objection to this index number is that each price relative exercises an equal influence upon the index number, whereas some price relatives are economically more important than others. Another objection that is sometimes raised is that the simple arithmetic mean is not an appropriate type of average to use with ratios, since it results in too high a value and is, therefore, said to have an "upward bias." The geometric mean is considered by many statisticians to be better. A *simple average of quantity relatives* can also be computed by use of

$$Q_{0j} = \frac{1}{n} \sum \left( \frac{q_j}{q_0} \right) \quad (18-4)$$

and forms a crude quantity index. Again, the same objections apply to this index as were noted for Eq. (18-3).

**TABLE 18.5: RELATED VEGETABLE OIL CONSUMPTION RELATIVE TO 1963, AND SIMPLE ARITHMETIC AVERAGE OF RELATIVES INDEX (PERCENT)**

<i>Type of oil</i>	1963	1964	1965	1966
	$\frac{q_0}{q_0}$	$\frac{q_1}{q_0}$	$\frac{q_2}{q_0}$	$\frac{q_3}{q_0}$
Soybean	100.0	114.3	114.0	134.5
Cottonseed	100.0	118.8	128.1	109.4
Linseed	100.0	96.9	59.4	59.4
Total	300.0	330.0	301.5	303.3
Index number*	100.0	110.0	100.5	101.1

\* Equation (18-4).

Source: Table 18.3.

## 18.5 AGGREGATIVE PRICE INDEX NUMBERS

It is often better to think of a price index number not as an average of relatives, frequently of heterogeneous data, but as a measure of the relative value in two different years of a fixed aggregate of goods. This type of price index number answers the question: "If we buy the same assortment of goods in each of two years, but at different prices, how much will we spend in the given year relative to the base year?" This is a particularly useful concept for consumer price indexes which compare the cost of supporting a particular plane of living in a given year with the cost in the base year.

**Weighted Aggregative Price Index Numbers.** In order to allow each commodity to have an appropriate influence on the index, it is advisable to use a weighted rather than a simple aggregate.<sup>(1)</sup> To construct a weighted aggregative index number, a list of definite quantities of specified commodities is taken, and calculations are made to determine what this aggregate of goods is worth in each of the two or more years under comparison. The value in any given year relative to the value in the base year is an aggregative price index number. Aggregative index numbers show merely the changing value of a fixed aggregate of goods. Since the total value changes while the components of the aggregate do not, the changes must be because of price.

<sup>(1)</sup> A simple aggregative index number is not, strictly speaking, unweighted. The weight assigned to each price is one unit of the commodity concerned.

There are innumerable possible aggregative price index numbers for a binary comparison, depending on the list of commodities and the number of units of each being compared. A few systems of weighting are discussed below and in some cases are illustrated.

1. *Base period quantities.* The formula for this method, sometimes known as Laspeyres' index number, is

$${}_0P_{0j} = \frac{\sum p_j q_0}{\sum p_0 q_0} \quad (18-5)$$

and is illustrated in Table 18.6. The zero subscript to the left of  $P_{0j}$  is used to stress the fact that base year weights are being employed. When used for a consumer price index number, this formula compares the theoretical cost in the given year with the actual cost in the base year of maintaining the standard of living of the base year. It has been noted, however, that the base period quantities may not be typical of those consumed in later years.

**TABLE 18.6: CONSTRUCTION OF AGGREGATIVE PRICE INDEX OF RELATED VEGETABLE OIL PRICES, USING BASE YEAR QUANTITY WEIGHTS (LASPEYRES' PRICE INDEX NUMBER)**

Type of oil	1963	1964	1965	1966
	$p_0 q_0$	$p_1 q_0$	$p_2 q_0$	$p_3 q_0$
Soybean	41.86	38.64	41.86	45.08
Cottonseed	14.40	13.44	14.40	17.28
Linseed	4.16	4.16	4.16	4.16
Total	60.42	56.24	60.42	66.52
Index number*	100.0	93.1	100.0	110.1

\* Equation (18-5).

Source: Tables 18.2 and 18.3.

2. *Given period quantities.* The formula for this method, sometimes known as Paasche's index number, is

$${}_jP_{0j} = \frac{\sum p_j q_j}{\sum p_0 q_j} \quad (18-6)$$

and is illustrated in Table 18.7. The  $j$  subscript to the left of  $P_{0j}$  stresses the fact that given year weights are being employed. When used for a consumer price index, this formula compares the actual cost in the given year with the theoretical cost in the base year of maintaining the plane of living of the given year. This method involves the selection of a new set of weights each year or even each month. Often it is impossible to obtain a revised set of weights so frequently. Even if this is possible, Table 18.7 clearly illustrates that the amount of computation required is nearly double that of the base year

TABLE 18.7: CONSTRUCTION OF AGGREGATIVE PRICE INDEX OF RELATED VEGETABLE OIL PRICES, USING GIVEN YEAR QUANTITY WEIGHTS (PAASCHE'S PRICE INDEX NUMBER)

Type of oil	1963		1963		1964		1964		1965		1965		1966	
	$P_{010}$	$P_{010}$	$\bar{P}_{010}$	$P_{011}$	$P_{111}$	$P_{111}$	$P_{012}$	$P_{112}$	$P_{013}$	$P_{113}$	$P_{014}$	$P_{114}$	$P_{015}$	$P_{115}$
Soybean	41.86	41.86	41.86	47.84	44.16	47.71	47.71	47.71	47.71	47.71	56.29	60.62	56.29	60.62
Cottonseed	14.40	14.40	14.40	17.10	15.96	18.45	18.45	18.45	18.45	18.45	15.75	18.90	15.75	18.90
Linseed	4.16	4.16	4.16	4.03	4.03	2.47	2.47	2.47	2.47	2.47	2.47	2.47	2.47	2.47
Total Index number*	60.42	60.42	60.42	68.97	64.15	68.63	68.63	68.63	68.63	68.63	74.51	81.99	74.51	81.99
			100.0		93.0							110.0		110.0

\* Equation (18-6).

Source: Tables 18.2 and 18.3.

weighted method.<sup>(9)</sup> A more serious objection is that given year weights may not be typical of earlier years. Furthermore, although each period is thereby directly compared with the base year, the comparison among the different years is not entirely valid for the reason that the quantities of the different goods change each year. It should be clear that both the Paasche and the Laspeyres price index numbers are open to serious criticism.

The Paasche index number is sometimes said to represent the lower limit of the price change, and the Laspeyres index number is said to represent the upper limit of the price change. If the demand schedules of consumers are assumed to be fixed, consumers will tend to purchase relatively more of commodities that have declined in price relative to other commodities and relatively less of commodities that have increased in price relative to other commodities. This rational action on the part of consumers will result in the Paasche price index number's being smaller than the Laspeyres price index number, since the Laspeyres price index number uses base period quantity weights. It is also sometimes said that the Laspeyres index number has an upward bias. The reason the word "bias" is used is that although the total values being compared are of market baskets that are fixed in a physical sense, they are varying with respect to the level of living represented. The Laspeyres price index number does not take into account the fact that over time individuals tend to substitute commodities that decline relatively in price for those that increase relatively in price. But a similar kind of statement can be made for the Paasche price index number. Consider a comparison between two geographical areas for the same time period. If one desired to formulate a clothing price index which compared the price of clothing in Samoa with that of Siberia, a "biased" result would be obtained by using as fixed weights either the number of sarongs consumed in Samoa or the number of overcoats consumed in Siberia.

The buying habits of people change also because of changes in demand schedules. These demand schedules change because of changes in real income, changes in environment, changes in taste, advertising, and for other reasons. If there are no changes in the prices of commodities, other than those brought about by changes in demand, the Paasche index number can be larger than the Laspeyres index number, and the remarks concerning limits and bias are exactly the opposite of those made in the preceding paragraph. On the other hand, some statisticians maintain that it is improper to try to measure changes in the price level when there are changes in demand. The writers are of the opinion that the term "bias" in index numbers should be avoided, since it implies that there is a true price index number, usually

---

<sup>(9)</sup> For these reasons most price indexes computed in the United States are variations of the Laspeyres index. The wholesale price index and the consumer price index are two important examples. For the consumer price index, weights are sometimes revised without actually changing the base of the index.



considered to be somewhere between the Laspeyres and the Paasche index number.<sup>(9)</sup> Although the search for a true index number is important in the study of the history of index number construction, the concept has been abandoned by many scholars. The important point is that we can never be assured that a fixed base index will reflect adequately changes in the cost of a fixed level of living, because of the possibility of substitution of commodities resulting from price changes.

3. *Average quantities of base and given year.* This compromise solution, sometimes called the Marshall-Edgeworth formula, is

$${}_M P_{0j} = \frac{\sum p_j(q_0 + q_j)}{\sum p_0(q_0 + q_j)} = \frac{\sum p_j \bar{q}_{0,j}}{\sum p_0 \bar{q}_{0,j}} \quad (18-7)$$

Again, as with the given year weights, we have shifting weights and a resultant lack of comparability among any pair of years except years 0 and  $j$ .

4. *Average quantities in several years.* This index number is an extension of Eq. (18-7) and is a solution that has been adopted often. If it is decided to use as weights the average quantity for the three years 1957, 1958, and 1959, the formula is

$${}_{1957-59} P_{0j} = \frac{\sum p_j \bar{q}_{57-59}}{\sum p_0 \bar{q}_{57-59}}$$

The weights used, however, will eventually become obsolete. When this is the case a new index can be constructed and spliced to the old one by using the method discussed in Sec. 18.7.

5. *Average of quantities for all the years included in the index.* Though perhaps an excellent solution for an historical study, this plan is impracticable if the index is to be kept up to date, since it means current revision of weights and continuous revision of the complete set of index numbers.

6. *Hypothetical quantities.* For example, the quantities might refer to the quantities of different commodities necessary to support some desirable standard of living.

7. *Geometric average of a pair of index numbers with different systems of weighting.* When the two index numbers are, respectively, those with base and given year weights, the resulting average is frequently referred to as Fisher's index number or the "ideal" index number.

$${}_F P_{0j} = \sqrt{\frac{\sum p_j q_0}{\sum p_0 q_0} \cdot \frac{\sum p_j q_j}{\sum p_0 q_j}} \quad (18-8)$$

or

$${}_F P_{0j} = \sqrt{{}_0 P_{0j} \cdot {}_j P_{0j}}$$

---

<sup>(9)</sup> Presumably a true price index number would compare the cost at two different times (or places) of obtaining the same level of *satisfaction*. For obvious reasons such an index number is virtually impossible to determine.

Although this method is said to have certain technical advantages,<sup>(4)</sup> it is difficult to say precisely just what it does measure.

Combining the results of Tables 18.6 and 18.7, we can easily compute Fisher's index numbers. (Note that in each year the arithmetic mean and the geometric mean of the two index numbers give the same results to one decimal place.)

<i>Index number</i>	<i>1963</i>	<i>1964</i>	<i>1965</i>	<i>1966</i>
Laspeyres	100.0	93.1	100.0	110.1
Paasche	100.0	93.0	100.0	110.0
Fisher	100.0	93.0	100.0	110.0

The alternatives mentioned may seem to be much ado about nothing, since the system of weights often makes little difference. Table 18.8 shows the results of using base year quantity weights that have been rounded to one digit. It is interesting to note that when the index numbers of Tables 18.6 and 18.7 are both rounded to the nearest percentage, the results are the same. (If the *prices* are rounded to one digit, however, the effect on the index numbers is considerable.)

**TABLE 18.8: CONSTRUCTION OF AGGREGATIVE INDEX OF RELATED VEGETABLE OIL PRICES, USING BASE YEAR WEIGHTS ROUNDED TO ONE SIGNIFICANT DIGIT**

<i>Type of oil</i>	<i>q<sub>0</sub></i>	<i>p<sub>0</sub>q<sub>0</sub></i>	<i>p<sub>1</sub>q<sub>0</sub></i>	<i>p<sub>2</sub>q<sub>0</sub></i>	<i>p<sub>3</sub>q<sub>0</sub></i>
Soybean	300	39	36	39	42
Cottonseed	100	15	14	15	18
Linseed	30	4	4	4	4
Total	430	58	54	58	64
Index number*		100.0	93.1	100.0	110.3

\* Equation (18.5).

Source: Tables 18.2 and 18.3.

But it so happens that if commodities that are changing *greatly* in relative importance during the period are also undergoing price changes materially different from the average, then the matter of weighting becomes important. Over a considerable period of time the changes in the prices and quantities are likely to be so great that any fixed base index number will become inaccurate. In recent years there has been a tendency to employ a type of index in which "link-relative" index numbers are chained together to form a

<sup>(4)</sup> See Sec. 18.8.

"chain" index. Each link relative is an index number with the preceding year as base and with weights appropriate for the two years under comparison. Chain index numbers will be considered in Sec. 18.7.

**Weighted Aggregative Quantity Index Numbers.** An aggregative index number of physical volume is the counterpart of the analogous aggregative price index number. Just as the aggregative index number of price measures the changing value of an aggregate of goods when the quantities have been held constant, so the aggregative index number of physical volume measures the changing value of a varying aggregate of goods when the prices have been held constant. The price index number answers the question: "If we buy the same assortment of goods in each of two years, but at *different prices*, how much will we spend in the given year relative to the base year?" The quantity index number answers the question: "If we buy *varying quantities* of specified goods in each of two years, but at the same price, how much will we spend in the given year relative to the base year?" Whereas in the former case the difference in the amount spent was attributable to price change, in the latter case the difference must, of course, be attributed to changes in quantities bought and sold, since prices were held constant. Three weighting systems are illustrated here:

1. Base year prices

$${}_0Q_{0j} = \frac{\sum p_0 q_j}{\sum p_0 q_0} \quad (18-9)$$

2. Given year prices

$${}_jQ_{0j} = \frac{\sum p_j q_j}{\sum p_j q_0} \quad (18-10)$$

3. Fisher's index number

$${}_FQ_{0j} = \sqrt{\frac{\sum p_0 q_j}{\sum p_0 q_0} \cdot \frac{\sum p_j q_j}{\sum p_j q_0}} \quad (18-11)$$

In general, the merits and defects of the different types of quantity index numbers are the same for the corresponding type of price index number. Their computation is left as an exercise (see Problem 1).

## 18.6 COMPUTATION BY THE METHOD OF WEIGHTED AVERAGE OF RELATIVES

Construction of index numbers by simple<sup>(5)</sup> averages of price relatives and quantity relatives was mentioned briefly in Sec. 18.4 and found to be unsatisfactory because of inappropriate weighting.

<sup>(5)</sup> A simple average-of-relatives index number is not, strictly speaking, unweighted. The value weight assigned to each relative is the same.

Weighted averages of relatives are obtained by multiplying each relative by its weight, summing these products, and dividing by the sum of the weights. The formulas thus are

$$P_{0j} = \frac{\sum \left[ v \left( \frac{p_j}{p_0} \right) \right]}{\sum v} \quad (18-12)$$

$$Q_{0j} = \frac{\sum \left[ v \left( \frac{q_j}{q_0} \right) \right]}{\sum v} \quad (18-13)$$

The weights are always values, the value weight for any commodity being the product of a price and a quantity ( $v = pq$ ). For a price index number the value is the product of the base year price and whatever quantity has been decided upon. For the quantity index number the value weight is the product of the base year quantity and whatever price has been decided upon.

Exactly the same results can be obtained by using aggregate values as by averaging relatives. Two systems of weighting averages of price relatives and two systems of weighting quantity relatives, together with their equivalent aggregative types, are as follows:

<i>Price index number</i>	<i>Average of relatives formula</i>	<i>Equivalent of aggregate formula</i>	
${}_0P_{0j}$ :	$\frac{\sum \left[ p_0 q_0 \left( \frac{p_j}{p_0} \right) \right]}{\sum p_0 q_0}$	$\frac{\sum p_j q_0}{\sum p_0 q_0}$	(18-14)

${}_jP_{0j}$ :	$\frac{\sum \left[ p_0 q_j \left( \frac{p_j}{p_0} \right) \right]}{\sum p_0 q_j}$	$\frac{\sum p_j q_j}{\sum p_0 q_j}$	(18-15)
----------------	-----------------------------------------------------------------------------------	-------------------------------------	---------

<i>Quantity index number</i>	<i>Average of relatives formula</i>	<i>Equivalent aggregate formula</i>	
${}_0Q_{0j}$ :	$\frac{\sum \left[ p_0 q_0 \left( \frac{q_j}{q_0} \right) \right]}{\sum p_0 q_0}$	$\frac{\sum p_0 q_j}{\sum p_0 q_0}$	(18-16)

${}_jQ_{0j}$ :	$\frac{\sum \left[ p_j q_0 \left( \frac{q_j}{q_0} \right) \right]}{\sum p_j q_0}$	$\frac{\sum p_j q_j}{\sum p_j q_0}$	(18-17)
----------------	-----------------------------------------------------------------------------------	-------------------------------------	---------

We will provide here only one illustration of a price index constructed by using a weighted average of relatives, an average of price relatives weighted with base year values. The construction of such an index is shown in Table 18.9. Notice that the last three columns of this table agree with the last three columns of Table 18.6.

**TABLE 18.9: CONSTRUCTION OF AVERAGE OF RELATIVES PRICE INDEX OF RELATED VEGETABLE OIL PRODUCTS, WEIGHTED WITH BASE YEAR WEIGHTS (1963 = 100)**

Type of oil	VALUE	PRICE RELATIVE			WEIGHTED PRICE RELATIVE		
	1963	1964	1965	1966	1964	1965	1966
	$v_0 = \frac{p_0 q_0}{p_0 q_0}$	$\frac{p_1}{p_0}$	$\frac{p_2}{p_0}$	$\frac{p_3}{p_0}$	$v_0 \left( \frac{p_1}{p_0} \right)$	$v_0 \left( \frac{p_2}{p_0} \right)$	$v_0 \left( \frac{p_3}{p_0} \right)$
Soybean	41.86	0.923	1.000	1.077	38.64	41.86	45.08
Cottonseed	14.40	0.933	1.000	1.200	13.44	14.40	17.28
Linseed	4.16	1.000	1.000	1.000	4.16	4.16	4.16
Total	60.42	...	...	...	56.24	60.42	66.52
Index number*	100.0	...	...	...	93.1	100.0	110.1

\* Equation (18-14).

Source: Price relatives from Table 18.4; value weights from Table 18.6.

There is a saving of labor if the weights are made to total unity, as in Table 18.10.

$${}_0P_{0j} = \Sigma \left[ w'_0 \left( \frac{p_j}{p_0} \right) \right] \quad (18-18)$$

where

$$w'_0 = \frac{p_0 q_0}{\Sigma p_0 q_0}$$

The resulting figures are, of course, the same as in Table 18.9. This refinement also permits us to see how many points each commodity contributed to the index number each year. Thus we see that in 1966 soybean oil contributed 75 out of 110 percentage points to the index number.

**TABLE 18.10: CONSTRUCTION OF AVERAGE OF RELATIVES INDEX OF RELATED VEGETABLE OIL PRODUCTS, WITH WEIGHTS PROPORTIONATE TO BASE YEAR TOTALING TO UNITY (1963 = 100)**

Type of oil	Proportionate value in base year $w'_0$	PRICE RELATIVE			WEIGHTED PRICE RELATIVE		
		1964	1965	1966	1964	1965	1966
		$\frac{p_1}{p_0}$	$\frac{p_2}{p_0}$	$\frac{p_3}{p_0}$	$w'_0 \left( \frac{p_1}{p_0} \right)$	$w'_0 \left( \frac{p_2}{p_0} \right)$	$w'_0 \left( \frac{p_3}{p_0} \right)$
Soybean	0.69282	0.923	1.000	1.077	0.63947	0.69282	0.74617
Cottonseed	0.23833	0.933	1.000	1.200	0.22236	0.23833	0.28600
Linseed	0.06885	1.000	1.000	1.000	0.06885	0.06885	0.06885
Total	1.00000	...	...	...	0.93068	1.00000	1.10102
Index number*	100.0	...	...	...	93.1	100.0	110.1

\* Equation (18-18).

Source: Table 18.9.

Various situations in which it is necessary or advantageous to use the average of relatives method of index number construction should be mentioned.

1. When a commodity is used to represent a group of commodities, its price relative is weighted by the value of the group. The alternative is to use a fictitious quantity weight obtained by dividing the value of the group of commodities by the price of the group representative. This is at best a confusing procedure.

2. When a commodity or series is to be substituted for one formerly used, the relative for the new commodity may be spliced to the relative for the old commodity. Former value weights are used.

3. The individual price or quantity relatives may be worth studying. The individual relatives having been computed, it is very simple to utilize them in constructing the index.

4. An index may be constructed by combining several previously constructed indexes, all of which have the same base.

5. When an index of cyclical fluctuations is to be constructed, it is necessary to adjust each original series for trend and seasonal variation. A discussion of the analysis of economic time series is begun in the next chapter. For the purpose of future reference we note here that usually the adjustment is made by dividing the time series by the estimated trend-seasonal, obtaining cyclical-irregular *relatives*. The different series are thus also put in comparable form.

6. Indexes of plant efficiency, level of living, and the like may be based on raw data that are not analogous to prices or quantities. Such data may take various forms, such as ratio of defectives to output, cost of production, median number of years of schooling, and so on. Such series can be made comparable by expressing them as ratios to (or percentages of) some base.

## 18.7 CHANGING THE BASE OF AN INDEX

It may be desirable to change the base of an existing index for various reasons, including: (1) to make the base more recent, (2) to permit easy comparison with some date of special interest (such as entry of the United States into World War II), (3) to provide a better comparison with some other index or series of relatives which has a different base, (4) to splice two overlapping indexes together, (5) to construct a chain index.

**Shifting to a More Recent Base.** As an illustration of this procedure, let us shift the base of the consumer price index shown in Table 18.11 from its 1957–1959 base to a 1960 base. The shifting is accomplished by

the simple device of dividing each of the index numbers of the original index by the index number for the new base year, i.e., 1.031.

**TABLE 18.11: CHANGING BASE OF CONSUMER PRICE INDEX FROM 1957-1959 = 100 TO 1960 = 100**

Year	INDEX NUMBER	
	1957-1959 = 100	1960 = 100 (1957-1959 base index numbers divided by 1.031)
1957	98.0	95.1
1958	100.7	97.7
1959	101.5	98.4
1960	103.1	100.0
1961	104.2	101.1
1962	105.4	102.2
1963	106.7	103.5
1964	108.1	104.8
1965	109.9	106.6
1966	113.1	109.7

Source: Board of Governors of the Federal Reserve System, Federal Reserve Bulletin, July 1967, p. 1226.

**Splicing Two Overlapping Indexes.** In Table 18.12 is given an aggregative index number (index A) which uses 1963 as a base with 1963 quantity weights. Now let us assume that we consider the 1963 weights to be outmoded and that we want 1964 weights. This second index, index B, is computed in the same manner as index A except for the use of 1964 weights. There is one overlapping year, 1964, for the two indexes. We wish to splice these two indexes together to form a continuous series. This operation is illustrated in Table 18.12. In the fourth column both indexes are put on a 1963 basis by multiplying index B by the 1964 value for index A, i.e., 0.931. In the last column both indexes are put on a 1964 basis by dividing index A by the 1964 value for index A, i.e., 0.931.

**TABLE 18.12: SPLICING TWO INDEX NUMBERS**

Year	Index A 1963 = 100	Index B 1964 = 100	SPLICED INDEX	
			1963 = 100 (Index B) · (0.931)	1964 = 100 (Index A)/0.931
1963	100.0	...	100.0	107.4
1964	93.1	100.0	93.1	100.0
1965	...	107.5	100.1	107.5
1966	...	118.6	110.4	118.6

$\text{Index A} = \frac{\sum p_1 q_{13}}{\sum p_{13} q_{13}}$  as given in Table 18.6.

$\text{Index B} = \frac{\sum p_1 q_{14}}{\sum p_{14} q_{14}}$ .

A procedure equivalent to that of Table 18.12 is followed whenever an index is revised. Such a revision usually involves not only changing the base and the weights, but also changing the commodities as well.<sup>(6)</sup>

**Chain Index.** Sometimes the list of commodities (or their specifications) and the system of weights are revised each year. The procedure is similar to that just explained. A series of link-relative index numbers is constructed in which each member is expressed as a percentage of the preceding year. This procedure yields a set of figures showing year-to-year comparisons, and it is in such terms that the businessman often thinks. These link relatives, if desired, may then be chained back to a fixed base by *successive* multiplication. Since the object of the chain index is to obtain maximum year-to-year comparability, the weighting system should be strictly up to date. Possibly the best solution is to use the ideal index number formula for each link relative.

$${}_F P_{j-1,j} = \sqrt{\frac{\sum p_j q_{j-1}}{\sum p_{j-1} q_{j-1}} \cdot \frac{\sum p_j q_j}{\sum p_{j-1} q_j}}$$

A solution that gets almost the same result is to use as quantity weights the average quantities of the current and preceding year.

$${}_M P_{j-1,j} = \frac{\sum p_j (q_{j-1} + q_j)}{\sum p_{j-1} (q_{j-1} + q_j)}$$

The student is asked to calculate this latter solution in Problem 2.

To illustrate the chaining of the link relatives back to a fixed base by successive multiplication, suppose that we had calculated the link relatives  ${}_M P_{01}$ ,  ${}_M P_{12}$ , and  ${}_M P_{23}$ . Then, if it were desired to have an index number for year three with year zero as the base, such an index number could be obtained by calculating

$${}_M P_{02} = {}_M P_{01} \cdot {}_M P_{12}$$

and

$${}_M P_{03} = {}_M P_{02} \cdot {}_M P_{23} = {}_M P_{01} \cdot {}_M P_{12} \cdot {}_M P_{23}$$

Although maximum comparability between successive years is obtained by a chain index, only the year following the fixed base is strictly comparable with that base. With the passage of time, many link relatives are involved, and the meaning of the chain index becomes increasingly doubtful.

## 18.8 TESTS OF INDEX NUMBERS

There are two sometimes conflicting purposes in constructing index numbers. One purpose is to find the answer to a specific question. For

<sup>(6)</sup> Changing the list of commodities is, of course, a special case of changing weights. When a commodity is dropped, its weight is changed from a positive quantity to zero; when a commodity is added, its weight is changed from zero to a positive quantity.



instance, one question might be: "What is the cost this year as compared with 1965 of supporting this year's scale of living?" Another question might be: "What is the cost this year as compared with 1965 of supporting the actual scale of living enjoyed each year, but at 1965 prices?" (It is suggested that the student write down the formula for the index numbers that will answer each of these questions.) The appropriate question to ask is a matter of economics or business administration, rather than of statistics.

But often an index is to be used by many persons for many purposes. Thus a general-purpose index number is wanted, one that will answer many questions approximately, but no question exactly.<sup>(7)</sup> The formula for such an index number is to be judged, in the opinion of some authorities, by whether it meets certain mathematical tests. The apparently reasonable assumption is made that since an index number is computed from a group of commodities, the index number should behave in the same manner as any individual commodity. For instance, for any individual commodity

$$\frac{p_1}{p_0} \cdot \frac{p_0}{p_1} = 1$$

$$\frac{p_1}{p_0} \cdot \frac{q_1}{q_0} = \frac{p_1 q_1}{p_0 q_0} = v, \text{ or value relative}$$

Therefore, two analogous tests, the time reversal test and the factor reversal test, have been laid down.

**Time Reversal Test.** This test holds that an index number should work backward as well as forward; an index number for position 0 relative to position  $j$  (the backward index number) should be the *reciprocal* of the index number for position  $j$  relative to position 0 (the forward index number). In general,

$$P_{0j} \cdot P_{j0} = 1$$

The backward index number formula is derived from the forward index number formula by reversing the time subscripts (0 and  $j$ ) of the forward index number formula. Thus, if  $\sum p_j q_0 / \sum p_0 q_0$  is the forward index number,  $\sum p_0 q_j / \sum p_j q_j$  is the backward index number. Obviously,

$$\frac{\sum p_j q_0}{\sum p_0 q_0} \cdot \frac{\sum p_0 q_j}{\sum p_j q_j} \neq 1$$

Therefore, Laspeyres' price index does not meet the time reversal test.<sup>(8)</sup> Also,

<sup>(7)</sup> Some people would say that the following is a general question: "How much has the price level increased since 1965?" Others regard it as a specific question.

<sup>(8)</sup> The "circular" test is an extension of the time reversal test. It says that if

$$P_{01} \cdot P_{12} \cdot \dots \cdot P_{j-1,j} \cdot P_{j0} = 1,$$

the same method of construction being used for each link relative, then the circular test is met. Most people who believe in the time reversal do not, however, believe in the circular test.

the student can verify that Paasche's price index does not meet the time reversal test.

**Factor Reversal Test.** This test holds that the product of a price index number and the corresponding quantity index number should accurately measure relative total value.

$$P_{0j} \cdot Q_{0j} = V_{0j} = \frac{\sum p_j q_j}{\sum p_0 q_0}$$

When a price index is likely to be used as a deflator, which obtains  $Q_{0j} = V_{0j}/P_{0j}$ , a plausible argument can be made for this test. A quantity index number formula corresponding to a price index number formula is derived from the latter by interchanging the factors ( $p$  and  $q$ ). Thus, if  $\sum p_j q_0 / \sum p_0 q_0$  is the price index number,

$$\frac{\sum q_j p_0}{\sum q_0 p_0} = \frac{\sum p_0 q_j}{\sum p_0 q_0}$$

is the quantity index number. Obviously, neither  ${}_0P_{0j}$  nor  ${}_jP_{0j}$  meets the factor reversal test.

There are several index number formulas that meet both tests, but only one is simple enough to justify serious consideration. This is the "ideal" index number, of which Irving Fisher was the leading advocate. As already stated,

$${}_F P_{0j} = \sqrt{\frac{\sum p_j q_0}{\sum p_0 q_0} \cdot \frac{\sum p_j q_j}{\sum p_0 q_j}}$$

If the factors are reversed, we obtain the ideal quantity index number.

$${}_F Q_{0j} = \sqrt{\frac{\sum p_0 q_j}{\sum p_0 q_0} \cdot \frac{\sum p_j q_j}{\sum p_j q_0}}$$

It is apparent that the product  ${}_F P_{0j}$  and  ${}_F Q_{0j}$  is  $\sum p_j q_j / \sum p_0 q_0$  and that the ideal index number meets the factor reversal test. It can be shown similarly that the ideal index number meets the time reversal test (see Problem 4).

The chief objections to the ideal index are (1) it is laborious to compute; (2) current quantity data are often difficult to obtain; (3) it seems impossible to say specifically what it measures, other than to say that it is the geometric mean of the Laspeyres and Paasche index numbers.

**Proportionality Test.** This test asserts that if each current year price in a price index is multiplied by a constant  $k$ , the resulting price index should be  $k$  times as large as before. It is easy to show that the Laspeyres, the Paasche and the Fisher index all meet this test.

## PROBLEMS

1. Using the data of Tables 18.2 and 18.3, calculate  ${}_0Q_{0j}$ ,  ${}_jQ_{0j}$ , and  ${}_FQ_{0j}$ . Explain in words what  ${}_0Q_{0j}$  and  ${}_jQ_{0j}$  mean.

2. Calculate link relatives by using  ${}_MP_{j-1,j}$  for the data of Tables 18.2 and 18.3.

3. Write down the formulas for index numbers that will answer questions posed in the first paragraph of Sec. 18.8.

4. Show that the Fisher price index number meets the time reversal test and the proportionality test.

5. Stuvell's index. Let  $P_{0j}$  and  $Q_{0j}$  be index numbers such that

$$P_{0j} \cdot Q_{0j} = V_{0j} = \frac{\sum p_j q_j}{\sum p_0 q_0}$$

Also, let the difference between Laspeyres' price index and  $P_{0j}$  equal the difference between Laspeyres' quantity index and  $Q_{0j}$ . Thus

$${}_0P_{0j} - P_{0j} = {}_0Q_{0j} - Q_{0j}$$

Showing that the resulting price index is

$$P_{0j} = \frac{{}_0P_{0j} - {}_0Q_{0j} + \sqrt{({}_0P_{0j} - {}_0Q_{0j})^2 + 4V_{0j}}}{2}$$

and that it does not meet the proportionality test.

6. A question often asked is: "Would you rather buy from a 1949 Sears catalogue or a 1969 Sears catalogue?" If most people answer 1969, does this indicate that the CPI should be lower in 1969 than it was in 1949? Discuss.

# 19

## Time Series Analysis: The Secular Trend

In the remaining chapters of this text we will direct our attention to the analysis of time series. Although our emphasis will be upon the analysis of economic time series, time series analysis is of interest to workers in many disciplines not directly related to business and economics.

### 19.1 THE PROBLEM OF TIME SERIES ANALYSIS

A time series is a set of observations arranged chronologically. The observations are usually, but not always, taken at equal intervals of time and are real numbers. A business firm's monthly sales figures extending from 1950 to 1969 offer a typical example of an economic time series. The yearly price and quantity index numbers of the last chapter offer other examples.

Simply stated, the problem of time series analysis is that of decomposition of the time series into rational components. The reasons for the attempt at decomposition vary. It may be that a certain component of the economic time series distorts the appearance of the components of interest, and, therefore, that its removal might lead to an improved understanding of the economic meaning of the components of direct interest. On the other hand, one might be interested in a certain component of the time series for its own sake, so that it may be compared to a similar component in another time series, utilized in forecasting, or subjected to economic analysis.

Traditionally an economic or business time series is said to have four components:

1. Secular trend *T*.
2. Cycles *C*.
3. Seasonal movements *S*.
4. Irregular fluctuations *I*.

Ordinarily we think of the time series *Y* as being the product of these components.<sup>(1)</sup>

$$Y = TCSI \quad (19-1)$$

The next two chapters will, in the main, be concerned with some techniques for the decomposition of the time series *Y* into the four components given in Eq. (19-1).

In this chapter we shall consider only the secular trend, the gradual increase or decrease over a period of time that is long relative to the other components.

There are several reasons for trend measurement. One reason is that it aids directly in business planning. If analysis of the physical volume of sales over a period of many years indicates that it has tended to grow in a particular way, it is fair to assume that it will continue to grow in that manner until there is a fundamental change in the system of causes in operation. The growth may be by a constant amount each year, or by a constant percentage, or in some other manner describable by a mathematical formula. Once the trend formula has been discovered, it is a simple matter to extend the trend any desired distance into the future. Having done so, we provide a basis for planning the financing and for constructing plant and equipment for future needs. It must be emphasized, however, (1) that projection of a trend is valid only if the correct trend type is selected, and only so long as the underlying system of causes governing the growth remains the same, and (2) that the trend equation is subject to error, and that the amount of error in the trend values becomes larger, sometimes very rapidly, the farther the trend is extended beyond the data.

A second reason for being interested in the trend is to help one in studying the other movements, especially the cycles, that fluctuate around the trend. In order to isolate the cycles, the trend values must be computed and then statistically eliminated from the data. This isolation is accomplished by dividing the original data by the trend values. If monthly or quarterly data are used, they must also be adjusted for seasonal variation. When data have

---

<sup>(1)</sup> Sometimes the model is taken to be additive

$$Y = T + C + S + I$$

but our attention will be directed toward the multiplicative model of Eq. (19-1). The multiplicative model may be made additive by writing

$$\log Y = \log T + \log C + \log S + \log I$$

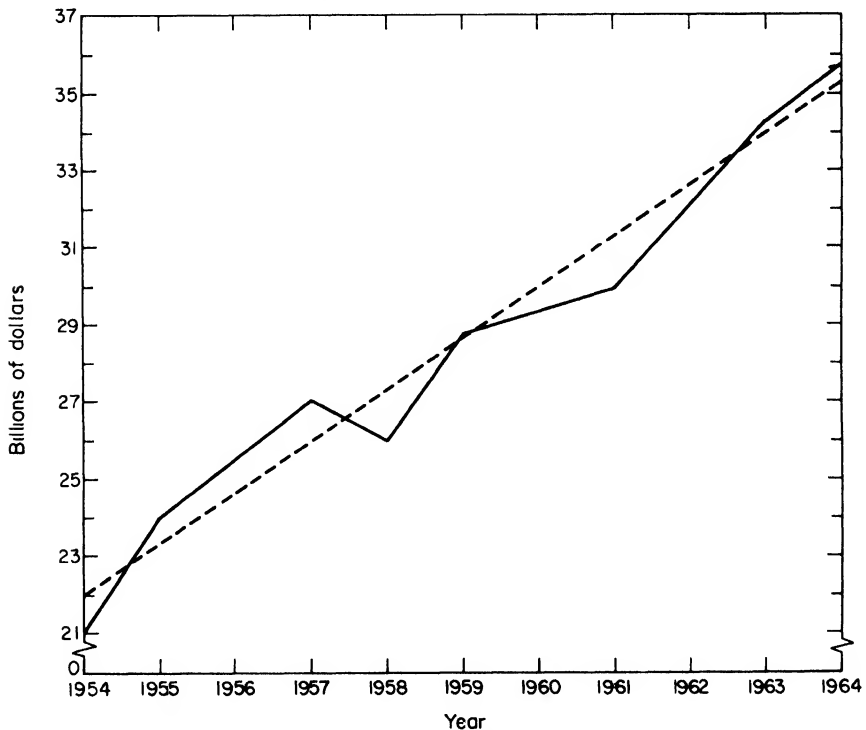
been adjusted for trend (or for seasonal and trend, if monthly or quarterly data are used), they may then be studied in order to make economic generalizations concerning cycles in the business or industry in question, and they may be compared with other similarly adjusted data to aid one in making forecasts.

A final reason for trend measurement is so that the rate or type of growth of two or more series may be compared. For example, we might find that the consumption of a certain commodity is growing at 5 percent per year in the Southern states and at only 3 percent per year in the New England states.

## 19.2 SOME EXAMPLES OF ECONOMIC TIME SERIES AND SECULAR TRENDS

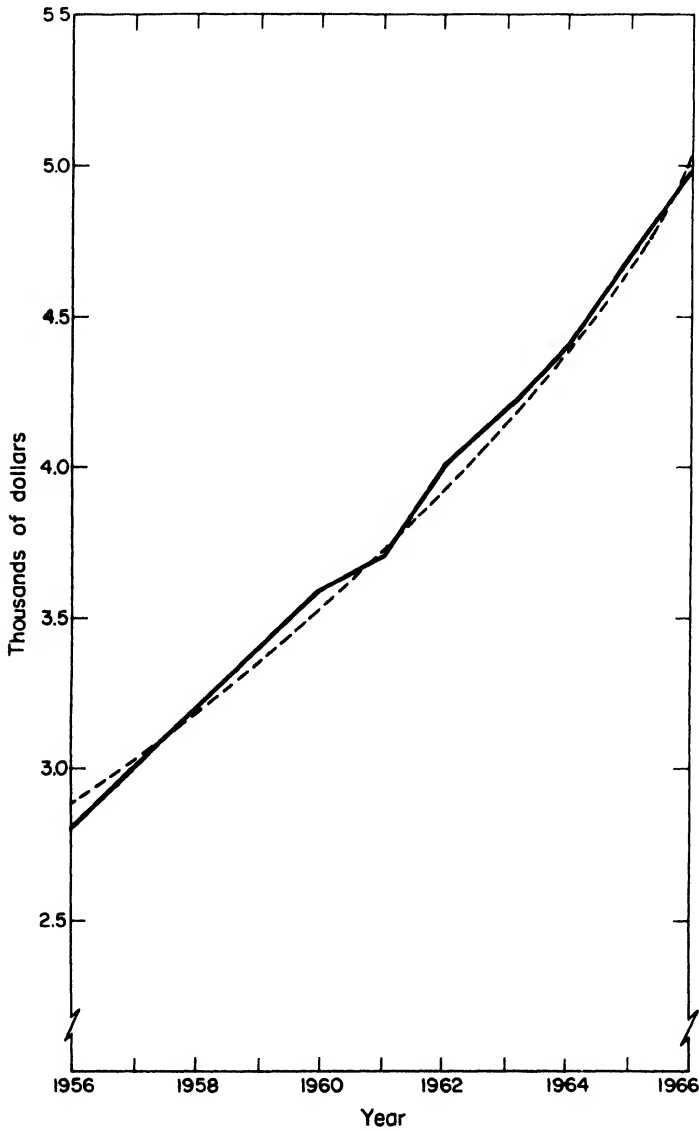
Chart 19.1 shows the annual sales made by the transportation industry in the United States, 1954–1964, expressed in billions of current dollars. The

**CHART 19.1: ANNUAL SALES BY THE TRANSPORTATION INDUSTRY AND LINEAR TREND, 1954–1964, BILLIONS OF CURRENT DOLLARS.**



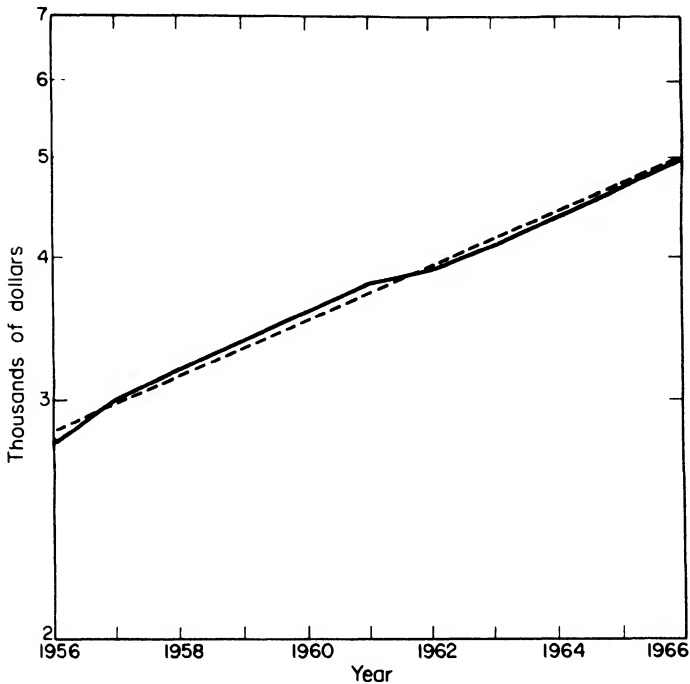
*Source: Department of Commerce, The National Income and Product Accounts of the United States, 1929–1965, Statistical Tables, Washington, 1966, p. 145.*

**CHART 19.2: AVERAGE SIZE OF ORDINARY LIFE INSURANCE POLICY AND EXPONENTIAL TREND, 1956-1966, THOUSANDS OF CURRENT DOLLARS (ARITHMETIC SCALE).**



*Source: Institute of Life Insurance, Life Insurance Fact Book, 1967, New York, p. 22.*

**CHART 19.3: AVERAGE SIZE OF ORDINARY LIFE INSURANCE POLICY AND EXPONENTIAL TREND, 1956-1966, THOUSANDS OF CURRENT DOLLARS (LOGARITHMIC SCALE).**



Source: Chart 19.2.

dotted line is the trend line. The fluctuations around the trend line appear to be mainly of a cyclical rather than of a random character; points above the trend line tend to come in groups as do points below the trend line.

Charts 19.2 and 19.3 show the average size of ordinary life insurance policies in the United States, 1956-1966, expressed in thousands of current dollars. Notice that on the arithmetic scale this time series appears to be nonlinear (it has an upward bend). However, on a chart that uses a logarithmic vertical scale, the time series appears to be linear. This time series is, in fact, a good example of a series with an approximately constant *percentage* change over time as contrasted with the transportation sales series, which serves to show a constant *amount* of change over time. A method of fitting the trend lines to these two time series will be illustrated in this chapter.

Both of the time series presented in this section appear to have trends that may be represented by straight lines. In the first case the trend is linear with regard to the original data; in the second case the trend is linear with regard to the logarithms of the original data and is said to be exponential. In



interpreting Charts 19.1 and 19.3 two things must be remembered. First, the series are in terms of current rather than constant dollars. The trends might appear to be of a different shape if the price element were removed (by dividing by a price index). Second, the period of time to which the trends were fitted is relatively short. Nearly any trend will appear to be linear if we confine ourselves to a sufficiently short segment of it.

### 19.3 TEST FOR SIGNIFICANCE OF TREND

It is foolish to fit a trend to a time series unless the time series is correlated with time. Because of the cyclical component of most time series, it is not valid to compute the simple correlation coefficient  $r$  and test its significance. Instead, some nonparametric measure of correlation is more appropriate (see Sec. 15.8). Using the transportation industry sales series of the last section, let us calculate Spearman's  $r_s$  as a measure of the significance of the linear trend. The calculation is given in Table 19.1 and follows that illustrated in Sec. 15.8.

**TABLE 19.1: ANNUAL SALES BY THE TRANSPORTATION INDUSTRY, 1954-1964, AND CALCULATION OF SPEARMAN'S RANK CORRELATION COEFFICIENT**

<i>Year</i>	<i>Annual sales</i>	<i>Year number rank</i>	<i>Sales rank</i>	<i>Difference in rank D</i>	<i>D<sup>2</sup></i>
1954	21.2	1	1	0	0
1955	24.2	2	2	0	0
1956	25.7	3	3	0	0
1957	27.2	4	5	-1	1
1958	25.9	5	4	1	1
1959	28.7	6	6	0	0
1960	29.3	7	7	0	0
1961	29.9	8	8	0	0
1962	32.2	9	9	0	0
1963	34.5	10	10	0	0
1964	35.8	11	11	0	0
Total	...	...	...	...	2

*Source: See Chart 19.1.*

Then, Spearman's  $r_s$  is

$$r_s = 1 - \frac{6 \sum D^2}{n(n^2 - 1)} = 1 - \frac{6(2)}{11(120)} = +0.91$$

Using the method of testing  $r_s$  given in Sec. 15.8, we find that  $r_s$  is highly significant. This is not surprising. The existence of a linear trend is obvious from a casual inspection of Chart 19.1.

## 19.4 FITTING A LINEAR TREND BY THE METHOD OF LEAST SQUARES

Various methods are available for fitting trends, each of which uses a different criterion. For the most part we shall use the method of least squares, with which the student is already familiar. The theoretical superiority of this method is not obvious, since the fluctuations about the trend lines are usually cyclical, rather than completely random, and not usually distributed normally. Yet the method of least squares is a convenient one.<sup>(2)</sup>

If we call  $Y$  the time series in question and let  $X$  represent time, then the linear trend is the time series stated as a linear function of time  $\hat{Y}$ . The equation, it should be recalled, is of the form

$$\hat{Y} = a + bX$$

and the method of least squares minimizes

$$\sum (Y - \hat{Y})^2$$

and leads to the normal equations

$$\left. \begin{aligned} na + b \sum X &= \sum Y \\ a \sum X + b \sum X^2 &= \sum XY \end{aligned} \right\} \quad (19-2)$$

which can be solved for

$$b = \frac{\sum xy}{\sum x^2} \quad (19-3)$$

and

$$a = \bar{Y} - b\bar{X} \quad (19-4)$$

Table 19.2 illustrates the calculation of the trend values shown on Chart 19.1. From Table 19.2 we see that

$$\begin{aligned} \sum xy &= \sum XY - \frac{(\sum X)(\sum Y)}{n} \\ &= 1715.5 - 1573 = 142.5 \\ \sum x^2 &= \sum X^2 - \frac{(\sum X)^2}{n} \\ &= 385 - 275 = 110 \end{aligned}$$

---

<sup>2</sup>Some other criteria used in this text in later sections are

1. The method of selected points (Sec. 19.8).
2. The method of partial totals (Sec. 19.8). Other criteria not illustrated are:
3. The method of maximum likelihood. Here, the best estimate of the population trend is taken to be the one that makes the observed sample most likely. For a straight line with normally distributed deviations, the estimate is the same as the least squares estimate.
4. The method of moments. The trend is fitted in such a way that  $\sum xY = \sum x\hat{Y}$ . For a straight line the estimate is the same as the least squares estimate.
5. Least absolute deviations. The trend equation is the one which minimizes  $\sum |Y - \hat{Y}|$ . This is a cumbersome method and is rarely used.

**TABLE 19.2: COMPUTATION OF LINEAR TREND FOR ANNUAL SALES BY TRANSPORTATION INDUSTRY, 1954-1964 (Y VALUES IN BILLIONS OF CURRENT DOLLARS)**

Year	Time $X$	Sales $Y$	$X^2$	$XY$	Trend value $\hat{Y}$
1955	0	21.2	0	0	22.1
1956	1	24.2	1	24.2	23.4
1957	2	25.7	4	51.4	24.7
1958	3	27.2	9	81.6	26.0
1959	4	25.9	16	103.6	27.3
1960	5	28.7	25	143.5	28.6
1961	6	29.3	36	175.8	29.9
1962	7	29.9	49	209.3	31.2
1963	8	32.2	64	257.6	32.5
1964	9	34.5	81	310.5	33.8
1965	10	35.8	100	358.0	35.1
Total	55*	314.6	385*	1715.5	314.6
Mean	5	28.6	...	...	...
Correction term	...	...	275	1573	...
Variation or covariation	...	...	110	142.5	...

\* The sums and sums of squares of the first 50 natural numbers are given in Appendix 15.

Source: See Chart 19.1.

so that 
$$b = \frac{142.5}{110} = 1.295$$

and 
$$a = \bar{Y} - b\bar{X}$$
$$= 28.6 - 1.295(5) = 22.1$$

The equation that estimates the linear trend is

$$\hat{Y} = 22.1 + 1.295X \quad (19-5)$$

and the trend values given in Table 19.2 are found by evaluating Eq. (19-5) for the given values of  $X$ .

Notice several things about Table 19.2. First,  $X$  is given the value of the integers zero through  $n - 1$ , where  $n$  is the number of observations. Actually,  $X$  can be any set of equally spaced numbers which is convenient. The year column itself could have been called  $X$ , but calculations would have been much more cumbersome. Second, Appendix 15 gives values of  $\sum X$  and  $\sum X^2$  so that it is not necessary to write down column four in Table 19.2. Third, the sum of the  $Y$  column is the same as the sum of the  $\hat{Y}$  column. This fact with which the student is already familiar offers a convenient check on the calculation.

### 19.5 MORE EFFICIENT CALCULATION OF THE LINEAR TREND BY THE METHOD OF LEAST SQUARES

In this section a calculation method for fitting a linear trend that is highly efficient is illustrated and explained. The method rests upon the fact that the values of the independent variables may be any set of evenly spaced integers.

Using the second normal equation given in Eqs. (19-2), let us write

$$a \sum x + b \sum x^2 = \sum xY$$

where  $x = X - \bar{X}$ . Then, since  $\sum x = 0$ , we see that

$$b = \frac{\sum xY}{\sum x^2}$$

From the first normal equation we write

$$na + b \sum x = \sum Y$$

so that

$$a = \frac{\sum Y}{n} = \bar{Y}$$

and it follows that

$$\hat{Y} = a + bx$$

**Odd Number of Observations.** The calculation of the trend estimates by the method of least squares for an odd number of observations is illustrated in Table 19.3, where a hypothetical time series is used. Notice that the middle  $x$  value is zero and that the other  $x$  values progress in units of one integer positively and negatively about this central value. Or, we may say that the  $x$  units are one year.

TABLE 19.3: EFFICIENT CALCULATION OF LINEAR TREND FOR A HYPOTHETICAL TIME SERIES (ODD NUMBER OF OBSERVATIONS)

<i>Year</i>	<i>X</i>	$x$ $X - \bar{X}$	<i>Time series</i> <i>Y</i>	$xY$	<i>Trend value*</i> $\hat{Y}$
1964	0	-2	10	-20	12
1965	1	-1	30	-30	21
1966	2	0	20	0	30
1967	3	1	40	40	39
1968	4	2	50	100	48
Sum	10	0	150	90	150
Mean	2	0	30	...	...
Sum of squares	...	10†	...	...	...

\* Calculated by using either  $\hat{Y} = 30 + 9x$  or  $\hat{Y} = 12 + 9X$ .

†  $\sum x^2$ .

Then, from Table 19.3

$$b = \frac{\sum xY}{\sum x^2} = \frac{90}{10} = 9$$

and

$$a = \bar{Y} = 30$$

so that the trend, in terms of  $x$  values, is

$$\hat{Y} = 30 + 9x \quad (19-6)$$

Notice that Eq. (19-6) has its origin at 1966 rather than 1964, since that is the year when  $x = 0$ . To put the origin of the trend equation at 1964, we notice that

$$x = X - \bar{X}$$

or, in this case,  $x = X - 2$ . Then, using Eq. (19-6), we replace  $x$  by  $X - 2$ .

$$\hat{Y} = 30 + 9(X - 2)$$

or

$$\hat{Y} = 12 + 9X \quad (19-7)$$

The trend values themselves will be unaffected by whether they are calculated by using Eq. (19-6) or Eq. (19-7).

**Even Number of Observations.** When the number of observations is even, it is convenient to use the transformation

$$x = 2(X - \bar{X}) \quad (19-8)$$

Table 19.4 uses the same  $Y$  values as did Table 19.3, with the exception of the addition of one  $Y$  value. The additional  $Y$  value lies directly on the regression line described by Eq. (19-7), so the end result of the calculation should be the same as before.

**TABLE 19.4: EFFICIENT CALCULATION OF LINEAR TREND FOR A HYPOTHETICAL TIME SERIES (EVEN NUMBER OF OBSERVATIONS)**

Year	$X$	$x$ $2(X - \bar{X})$	Time series $Y$	$xY$	Trend values* $\hat{Y}$
1964	0	-5	10	-50	12
1965	1	-3	30	-90	21
1966	2	-1	20	-20	30
1967	3	1	40	40	39
1968	4	3	50	150	48
1969	5	5	57	285	57
Sum	15	0	207	315	207
Mean	2.5	0	34.5	...	...
Sum of squares	...	70†	...	...	...

\* Calculated by using either  $\hat{Y} = 34.5 + 4.5x$  or  $\hat{Y} = 12 + 9X$ .

†  $\sum x^2$ .

Then, from Table 19.4

$$b = \frac{\sum xY}{\sum x^2} = \frac{315}{70} = 4.5$$

and

$$a = \bar{Y} = 34.5$$

so that the linear trend, in terms of  $x$  values, is

$$\hat{Y} = 34.5 + 4.5x \quad (19-9)$$

Notice that Eq. (19-9) has its origin midway between 1966 and 1967. Notice also that the  $b$  value of Eq. (19-9) is one-half as large as it was for Eqs. (19-6) and (19-7), because of the fact that the  $x$  values are spaced twice as far apart as before; the  $x$  units are one-half year. To express Eq. (19-9) in terms of  $X$  values we have only to use Eq. (19-9), replace  $x$  with  $2(X - \bar{X})$  and, since  $\bar{X} = 2.5$ , write

$$\hat{Y} = 34.5 + 4.5[2(X - 2.5)]$$

or

$$\hat{Y} = 12 + 9X \quad (19-10)$$

Equations (19-10) and (19-7) are found to be identical.

## 19.6 CHANGING UNITS AND SHIFTING ORIGIN

If the trend equation has been calculated from annual totals, i.e., the total of 12 monthly figures, or from monthly data given at annual rates, the trend equation may be expressed in terms of monthly averages by dividing  $a$  and  $b$ , respectively, by 12. This procedure is reasonable since, on the average, the monthly figures will be one-twelfth as large as the annual total. Thus, if

$$\hat{Y} = a + bX \quad (19-11)$$

is in terms of annual totals, then the equation in terms of monthly averages is

$$\hat{Y} = \frac{a}{12} + \frac{b}{12}X \quad (19-12)$$

Equation (19-12) has as its origin July 1 of the first year of the time series. The  $b$  coefficient in Eq. (19-12) shows the annual increase of the monthly average value of the trend. The  $X$  values are in units of one year.

If we wish to write a monthly trend equation where the  $X$  values are in units of one month and the  $b$  coefficient represents the estimate of the month-to-month increment in the trend, we transform Eq. (19-11) as follows:

$$\hat{Y} = \frac{a}{12} + \frac{b}{144}X \quad (19-13)$$

Equation (19-13), like Eq. (19-12), has July 1 of the first year as its origin. If we wish to have the middle of January of the first year as the origin,

Eq. (19-13) is rewritten as

$$\hat{Y} = \frac{a}{12} + \frac{b}{144} (X - 5.5) \quad (19-14)$$

The reason for subtracting 5.5 is that the middle of January of the first year is 5.5 months earlier than the middle of the first year, July 1.

## 19.7 HIGHER DEGREE POLYNOMIAL TRENDS

In the past sections we have discussed a polynomial trend of the first degree, i.e., a linear trend. In general, a polynomial trend has the form

$$\hat{Y} = a + bX + cX^2 + dX^3 + \dots$$

and the degree of the polynomial is given by the highest exponent to which  $X$  is raised. As the degree of the polynomial increases one degree, one new term is added to the equation and one new direction of slope is possible. For trend fitting experience seems to show that it is seldom desirable to use a polynomial higher than third degree.

**Second-degree Polynomial.** Chart 19.4 illustrates a second degree polynomial trend, also referred to as a quadratic trend or a parabola, given by the equation

$$Y = 10 + 5X + 2X^2 \quad (19-15)$$

Table 19.5 gives selected points at intervals of one  $X$  unit for Eq. (19-15) as well as the first and second differences of the  $Y$  values.<sup>(3)</sup> The first differences

**TABLE 19.5: SELECTED POINTS, FIRST AND SECOND DIFFERENCES OF  $Y$  VALUES GIVEN BY THE EQUATION  $Y = 10 + 5X + 2X^2$**

$X$	$Y$	$\Delta Y$	$\Delta^2 Y$
0	10		
1	17	7	
2	28	11	4
3	43	15	4
4	62	19	4

<sup>(3)</sup> The first difference  $\Delta Y_X$  is defined as

$$\Delta Y_X = Y_X - Y_{X-1}$$

and the second difference is the first difference of the first differences, *not* the square of the first difference

$$\Delta^2 Y_X = \Delta(\Delta Y_X)$$

and so on for higher differences.

represent the average slope between two points. The second differences have a constant value  $2c$ , which is 4 for our example. Two directions of slope are possible for a second-degree polynomial.<sup>(4)</sup> Finally, the second-degree polynomial need not be concave from above, as in our example, but, as we shall see, may be concave from below.

**CHART 19.4: GRAPH OF EQUATION  $Y = 10 + 5X + 2X^2$  (SECOND-DEGREE POLYNOMIAL).**

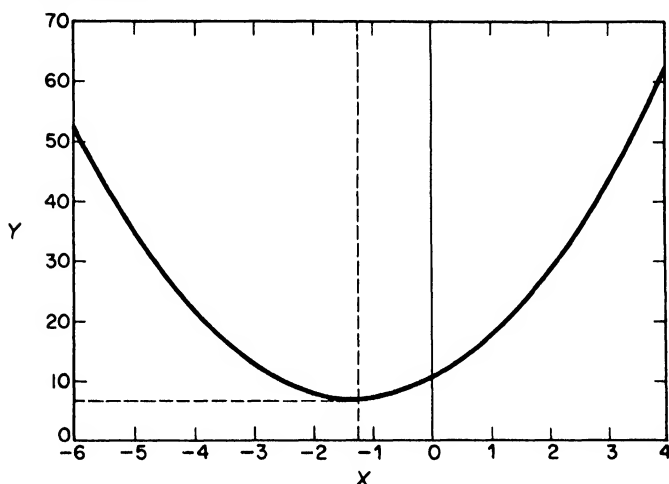


Chart 19.5 shows the percentage of assets of United States life insurance companies held in state, provincial, and local bonds, 1960–1966, and a second-degree polynomial trend that is indicated by the dotted line.

In order to fit a second-degree polynomial using the method of least squares, one must have three normal equations, since there are three constants. The normal equations are<sup>(5)</sup>

$$\left. \begin{aligned} na + b \sum X + c \sum X^2 &= \sum Y \\ a \sum X + b \sum X^2 + c \sum X^3 &= \sum XY \\ a \sum X^2 + b \sum X^3 + c \sum X^4 &= \sum X^2 Y \end{aligned} \right\} \quad (19-16)$$

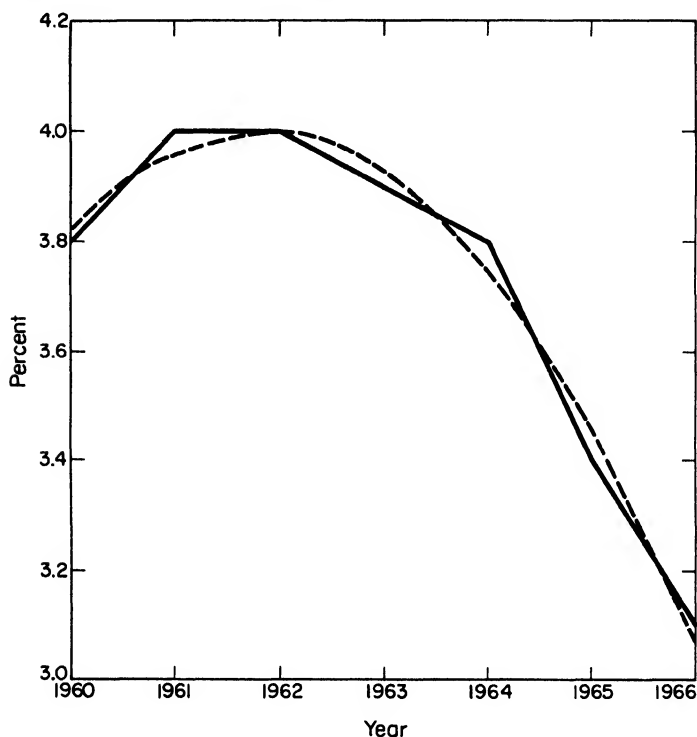
If, as usual, we express  $X$  in terms of deviations [ $x = X - \bar{X}$  for an odd number of observations and  $x = 2(X - \bar{X})$  for an even number of

<sup>(4)</sup> The student should verify that a first-degree polynomial (straight line) has constant first differences and only one possible direction of slope.

<sup>(5)</sup> See Sec. 16.2.



**CHART 19.5: PERCENTAGE OF ASSETS OF U.S. LIFE INSURANCE COMPANIES HELD IN STATE, PROVINCIAL, AND LOCAL BONDS, 1960-1966, AND SECOND-DEGREE POLYNOMIAL TREND.**



*Source: Institute of Life Insurance, Life Insurance Fact Book, 1967, New York, p. 63.*

observations], we may write Eq. (19-16) as

$$\left. \begin{aligned} na + c \sum x^2 &= \sum Y \\ b \sum x^2 &= \sum xY \\ a \sum x^2 + c \sum x^4 &= \sum x^2 Y \end{aligned} \right\} \quad (19-17)$$

since the sums of all odd powers of  $x$  vanish.

From the second normal equation in Eqs. (19-17)

$$b = \frac{\sum xY}{\sum x^2} \quad (19-18)$$

The first and last normal equations in Eqs. (19-17) must be solved simultaneously to obtain  $a$  and  $c$ . Note that if the equations  $\hat{Y} = a + bx$  and

**TABLE 19.6: CALCULATION OF SECOND-DEGREE POLYNOMIAL TREND FOR PERCENTAGE OF ASSETS OF U.S. LIFE INSURANCE COMPANIES HELD IN STATE, PROVINCIAL, AND LOCAL BONDS, 1960-1966**

<i>Year</i>	<i>X</i>	<i>x</i>	<i>x</i> <sup>2</sup>	<i>Y</i>	<i>xY</i>	<i>x</i> <sup>2</sup> <i>Y</i>	<i>Trend value*</i> <i>Ŷ</i>
1960	0	-3	9	3.8	-11.4	34.2	3.82
1961	1	-2	4	4.0	-8.0	16.0	3.96
1962	2	-1	1	4.0	-4.0	4.0	4.00
1963	3	0	0	3.9	0	0	3.93
1964	4	1	1	3.8	3.8	3.8	3.75
1965	5	2	4	3.4	6.8	13.6	3.46
1966	6	3	9	3.1	9.3	27.9	3.07
Total	21	0	28	26.0	-3.5	99.5	26.0
Mean	3	...	...	...	...	...	...
Sum of squares	...	...	196†	...	...	...	...

\* Calculated by using either  $\hat{Y} = 3.9286 - 0.125x - 0.5357x^2$  or  $\hat{Y} = 3.82147 + 0.19642X - 0.05357X^2$ .

†  $\sum x^4 = \sum (x^2)^2$ .

Source: See Chart 19.5.

$\hat{Y} = a + bx + cx^2$  are fitted to the same data, the values of  $b$  are the same, but the values of  $a$  are different.

Table 19.6 illustrates the necessary preliminary calculations, and from it we obtain

$$b = \frac{\sum xY}{\sum x^2} - \frac{-3.5}{28} = -0.125$$

and the first and third normal equations are

$$\begin{aligned} 7a + 28c &= 26 \\ 28a + 196c &= 99.5 \end{aligned}$$

Multiplying the first equation by  $\frac{2}{7} = 4$  and subtracting the third from the result, we find

$$\begin{aligned} -84c &= 4.5 \\ c &= -0.05357 \end{aligned}$$

so that

$$7a + 28(-0.05357) = 26$$

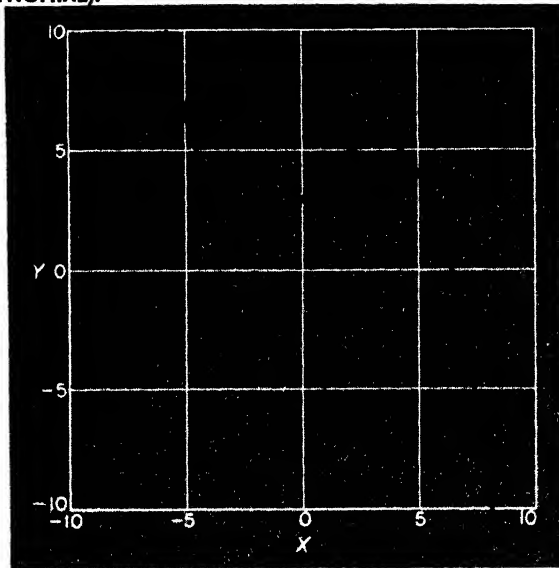
and

$$a = \frac{27.49996}{7} = 3.9286$$

The accuracy of the calculation of  $a$  and  $c$  may be checked by direct insertion in the normal equations.

The second-degree polynomial in terms of  $x$  values is

$$\hat{Y} = 3.9286 - 0.125x - 0.05357x^2$$

**CHART 19.6: GRAPH OF EQUATION  $Y = 3 + 1X + 0.1X^2 - 0.05X^3$  (THIRD-DEGREE POLYNOMIAL).**

In terms of  $X$  values, since  $\bar{X} = 3$ ,

$$\hat{Y} = 3.9286 - 0.125(X - 3) - 0.05357(X - 3)^2$$

$$\hat{Y} = 3.82147 + 0.19642X - 0.05357X^2$$

Again, we notice from Table 19.6 that the sum of the  $Y$  column is the same as the sum of the trend value column.

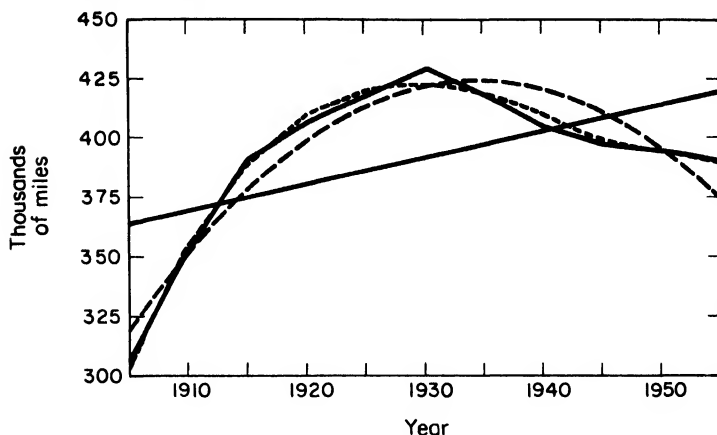
**Third-degree Polynomial.** Chart 19.6 illustrates a third-degree polynomial trend, also referred to as a cubic, given by the equation

$$Y = 3 + 1X + 0.1X^2 - 0.05X^3 \quad (19-19)$$

Table 19.7 gives selected points at intervals of one  $X$  unit for Eq. (19-19), as well as the first, second, and third differences of the  $Y$  values. The third differences have a constant value  $6d$ . A third-degree polynomial has three

**TABLE 19.7: SELECTED POINTS, FIRST, SECOND, AND THIRD DIFFERENCES OF  $Y$  VALUES GIVEN BY THE EQUATION  $Y = 3 + 1X + 0.1X^2 - 0.05X^3$** 

$X$	$Y$	$\Delta Y$	$\Delta^2 Y$	$\Delta^3 Y$
0	3.00	1.05		
1	4.05	0.95	-0.10	
2	5.00	0.55	-0.40	-0.30
3	5.55	-0.15	-0.70	-0.30
4	5.40			

**CHART 19.7: MILES OF RAILROAD TRACK OPERATED, AND POLYNOMIAL TRENDS, 1905-1955.**

Source: *Association of American Railroads, A Chronology of American Railroads, 1957, p. 8.*

possible directions of slope and one point of inflection, the point where the curve changes from concave from above to concave from below, or vice versa. On Chart 19.6 the point of inflection is reached where  $X = \frac{2}{3}$  and  $Y = 3.7$ .

Chart 19.7 shows the miles of railroad track in operation in the United States, 1905-1955, and a first-, second-, and third-degree polynomial trend. By looking at the chart, we can get an idea when the equation is of high enough degree to be a satisfactory representation of the trend. The third-degree trend appears to be the most satisfactory of the three.

The form of the third-degree estimating equation is

$$\hat{Y} = a + bX + cX^2 + dX^3$$

and the normal equations are

$$\left. \begin{aligned} na + b \sum X + c \sum X^2 + d \sum X^3 &= \sum Y \\ a \sum X + b \sum X^2 + c \sum X^3 + d \sum X^4 &= \sum XY \\ a \sum X^2 + b \sum X^3 + c \sum X^4 + d \sum X^5 &= \sum X^2Y \\ a \sum X^3 + b \sum X^4 + c \sum X^5 + d \sum X^6 &= \sum X^3Y \end{aligned} \right\} \quad (19-20)$$

The  $a$  and  $c$  constants in the estimating equation may be found by solving for  $a$  and  $c$  simultaneously in terms of  $x$ .

$$\left. \begin{aligned} na + c \sum x^2 &= \sum Y \\ a \sum x^3 + c \sum x^4 &= \sum x^3Y \end{aligned} \right\} \quad (19-21)$$

TABLE 19.8: CALCULATION OF THIRD-DEGREE POLYNOMIAL TREND FOR MILES OF RAILROAD TRACK OPERATED, 1905-1955

Year	X	x	x <sup>2</sup>	x <sup>3</sup>	x <sup>4</sup>	Y	XY	x <sup>2</sup> Y	x <sup>3</sup> Y	x <sup>4</sup> Y	Trend value* $\hat{Y}$
1905	0	-5	25	-125	625	307	-1535	7,675	-38,375	191,875	305.06
1910	1	-4	16	-64	256	352	-1408	5,632	-22,528	90,112	355.05
1915	2	-3	9	-27	81	391	-1173	3,519	-10,557	31,671	389.26
1920	3	-2	4	-8	16	407	-814	1,628	-3,256	6,512	410.17
1925	4	-1	1	-1	1	418	-418	418	-418	418	420.23
1930	5	0	0	0	0	430	0	0	0	0	421.90
1935	6	1	1	1	1	419	419	419	419	419	417.65
1940	7	2	4	8	16	406	812	1,624	3,248	6,496	409.93
1945	8	3	9	27	81	398	1194	3,582	10,746	32,238	401.21
1950	9	4	16	64	256	396	1584	6,336	25,344	101,376	393.94
1955	10	5	25	125	625	391	1955	9,775	48,875	244,375	390.60
Total	55	0	110	0	1958	4315	616	40,608	13,498	705,492	4315
Mean	5	...	...	...	...	...	...	...	...	...	...
Sum of squares	...	...	...	41,030†	...	...	...	...	...	...	...

\* Calculated by using either  $\hat{Y} = 421.899 - 1.699x - 2.9627x^2 + 0.4101x^3$  or  $\hat{Y} = 305.0640 + 58.6855X - 9.1142X^2 + 0.4101X^3$ .

†  $\sum x^4 = \sum (x^2)^2$ .

Source: See Chart 19.7.

just as was done for the second-degree polynomial. The constants  $b$  and  $d$  are found by solving

$$\left. \begin{aligned} b \sum x^2 + d \sum x^4 &= \sum xY \\ b \sum x^4 + d \sum x^6 &= \sum x^3Y \end{aligned} \right\} \quad (19-22)$$

Table 19.8 illustrates the calculations for the railroad data. From Table 19.8 and Eqs. (19-21) we find

$$\begin{aligned} 11a + 110c &= 4315 \\ 110a + 1958c &= 40,608 \end{aligned}$$

and solving the equations simultaneously in the usual way, we find

$$c = -2.9627$$

and

$$a = 421.899$$

Again, the calculated values of  $a$  and  $c$  may be checked by direct substitution into the normal equations directly above.

Next, using Eqs. (19-22), we find from Table 19.8 that

$$\begin{aligned} 110b + 1958d &= 616 \\ 1958b + 41,030d &= 13,498 \end{aligned}$$

Therefore,

$$d = 0.4101$$

and

$$b = -1.699$$

which may be checked by direct substitution.

The estimating equation, in terms of  $x$  values, is

$$\hat{Y} = 421.899 - 1.699x - 2.9627x^2 + 0.4101x^3$$

and in terms of  $X$  values is

$$\hat{Y} = 421.899 - 1.699(X - 5) - 2.9627(X - 5)^2 + 0.4101(X - 5)^3$$

or

$$\hat{Y} = 305.0640 + 58.6855X - 9.1142X^2 + 0.4101X^3$$

**Additional Comments.** Polynomial equations often fit the data set well within their range. However, it is usually impossible to find any logical basis for a polynomial equation. This fact will perhaps seem obvious when we note that a straight line or a third-degree polynomial trend will, upon extension, become negative in one direction and increase without limit on one end. On the other hand, most economic time series cannot have negative values and tend to reach an upper or lower limit in one or both directions.

A polynomial trend can be made to fit the observed data set as desired by the simple expedient of increasing the degree of the polynomial. If the degree of the polynomial is one less than the number of observations to which the polynomial is to be fit, the trend will coincide with every data point, but the trend will be meaningless. The meaninglessness of the trend arises in this

limiting case because one degree of freedom is lost every time a constant is added to the trend equation, i.e., every time the degree of the polynomial is increased by one. Thus, when the degree of the polynomial is one less than the number of observations, there are no degrees of freedom remaining. With an economic time series there are rarely  $n$  independent observations, since the  $Y$  values are autocorrelated; that is, since there is correlation between the values of  $Y$  at time  $X$ , time  $X + 1$ ,  $X + 2$ , etc. Because of this autocorrelation there is often autocorrelation in the  $(Y - \hat{Y})$  residuals. The effect of this autocorrelation is to reduce the number of degrees of freedom and to make the testing of the significance of the trend constants by the techniques outlined in previous chapters, inappropriate. Furthermore, the polynomial trend becomes meaningless well before the degree of the polynomial equals  $n - 1$ .

Calculation of higher degree polynomial trends becomes quite cumbersome unless one uses a digital computer or some other computational expedient. If a computer is available to the researcher, he can make use of any one of a number of programs that carry out polynomial regression.<sup>(6)</sup> If a desk calculator is all that is available, the labor of trend fitting can be reduced (except for straight line trends) by use of orthogonal polynomials.<sup>(7)</sup> An orthogonal polynomial equation is of the form

$$Y = \sum_{r=0}^m B_r \phi_r = B_0 \phi_0 + B_1 \phi_1 + B_2 \phi_2 + \cdots + B_m \phi_m$$

The  $\phi_r$  are orthogonal polynomials (uncorrelated polynomials). In the equation above,  $\phi_0 = 1$ ,  $\phi_1 = x$ ,  $\phi_2 = x^2 - \text{mean value of } x^2$ . The derivation of the other  $\phi_r$  values is more complicated, but the  $\phi_r$  values are uncorrelated with each other; i.e.,  $\sum_1^n \phi_r \phi_s = 0$ . Also,  $\sum \phi_r = 0$  ( $r \neq 0$ ). The  $B_r$  values are all computed by the formula

$$B_r = \frac{\sum \phi_r Y}{\sum \phi_r^2}$$

Since  $\phi_0 = 1$ ,  $B_0 = \bar{Y}$ ; since  $\phi_1 = x$ ,  $B_1 = \sum xY / \sum x^2$ . The values of  $B_r$  do not change as the degree of equation is increased. Use of orthogonal polynomials also facilitates appropriate tests of significance.

<sup>(6)</sup> Polynomial regression is, of course, simply a special case of multiple regression. Programs suitable for the IBM System/360, as well as for several other types of computers, which carry out polynomial regression, even including generation of powers of the independent variable, are described and explained in: *IBM System/360 Scientific Subroutine Package (360A-CM-03X) Programmer's Manual*, IBM Corp., 1966.

<sup>(7)</sup> Use of orthogonal polynomials is further explained and illustrated in F. E. Croxton and D. J. Cowden, *Practical Business Statistics*, 3rd ed. (Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1960), Chapter 34. See also E. S. Pearson and H. O. Hartley, *Biometrika Tables for Statisticians*, Vol. 1, 3rd ed., Cambridge University Press, 1966.

To fit a trend by use of orthogonal polynomials, a table of orthogonal polynomials for each  $n$  may be used. The  $\phi_r$  are always integers, and they usually contain only a small number of digits. Such tables provide not only the values of the individual polynomials  $\phi_r$ , but also  $\sum_{r=1}^n \phi_r^2$ .

## 19.8 GROWTH CURVES

A growth curve is here defined as one in which the amount of growth at any point of time (or during any period of time) is a function of the level attained at that point of time (or at the beginning of that period of time). In this section we shall consider these growth curves: exponential, second-degree exponential, modified exponential, Gompertz and logistic.

One procedure for fitting an exponential is to put the equation in linear form and then to use the method of least squares. For the Gompertz and logistic, the easiest procedure is to convert them into a modified exponential form and use the method of selected points or the method of partial totals.

**Exponential.** The exponential equation, which has the form

$$\hat{Y} = AB^x$$

describes a series that is changing by a constant ratio  $B$  per unit of time. Thus, if  $B = 1.10$  the trend value for each year is 110 percent of the previous year, and the rate of growth is 10 percent. The exponential curve is also called a compound interest curve, and on paper with a vertical logarithmic scale it is a straight line.

$$\log \hat{Y} = \log A + (\log B)x$$

or, if we let  $a = \log A$ ,  $b = \log B$

$$\log \hat{Y} = a + bx$$

In terms of  $x$  values

$$\log \hat{Y} = \bar{Y}_{\log} + bx$$

where  $\bar{Y}_{\log}$  is the arithmetic mean of the  $\log Y$  values.

We have already illustrated an example of an exponential curve in Charts 19.2 and 19.3. The data plotted on those charts is given in Table 19.9, which illustrates the fitting of an exponential curve. The calculation procedure is exactly the same as that for a straight line, except that the first step is to look up the logarithms of the  $Y$  values, and the last step is to look up the antilogarithms of the  $\log \hat{Y}$  values.

From Table 19.9 we find

$$b = \frac{\sum x \log Y}{\sum x^2} = \frac{2.60738}{110} = 0.0237035$$



**TABLE 19.9: CALCULATION OF EXPONENTIAL TREND FOR AVERAGE SIZE OF ORDINARY LIFE INSURANCE POLICIES, 1956-1966, THOUSANDS OF CURRENT DOLLARS**

Year	$X$	$x$ $X - \bar{X}$	$Y$	$\log Y$	$x \log Y$	$\log \hat{Y}$	Antilog ( $\log \hat{Y}$ ) trend value $\hat{Y}$
1956	0	-5	2.8	0.44716	-2.23580	0.45481	2.8
1957	1	-4	3.0	0.47712	-1.90848	0.47851	3.0
1958	2	-3	3.2	0.50515	-1.51545	0.50221	3.2
1959	3	-2	3.4	0.53148	-1.06296	0.52592	3.4
1960	4	-1	3.6	0.55630	-0.55630	0.54962	3.5
1961	5	0	3.8	0.57978	0	0.57333	3.7
1962	6	1	3.9	0.59106	0.59106	0.59703	3.9
1963	7	2	4.1	0.61278	1.22556	0.62073	4.2
1964	8	3	4.4	0.64345	1.93035	0.64444	4.4
1965	9	4	4.7	0.67210	2.68840	0.66814	4.7
1966	10	5	4.9	0.69020	3.45100	0.69184	4.9
Total	55	0	41.8	6.30658	2.60738	6.30658	41.7
Sum of squares	...	110*	...	...	...	...	...

\*  $\sum x^2$ .

Source: See Chart 19.2.

and 
$$\bar{Y}_{\log} = \frac{6.30658}{11} = 0.573325$$

Then, in logarithmic form and in terms of  $x$  values

$$\log \hat{Y} = 0.573325 + 0.0237035x$$

In terms of  $X$  values

$$\begin{aligned} \log \hat{Y} &= 0.573325 + 0.0237035(X - 5) \\ &= 0.454808 + 0.0237035X \end{aligned} \quad (19-23)$$

If we wish to take the antilog of each side of Eq. (19-23), we obtain

$$\hat{Y} = 2.848(1.056)^X \quad (19-24)$$

with  $X$  units of one year. Equation (19-24) indicates that the series is growing at almost 6 percent per year.

It is worth noting that  $\sum \log Y = \sum \log \hat{Y}$ , but  $\sum Y \neq \sum \hat{Y}$  (see Table 19.9). However, the geometric means of the  $Y$  and  $\hat{Y}$  values are equal. Also  $\sum (\log Y - \log \hat{Y})^2$  is a minimum, but not  $\sum (Y - \hat{Y})^2$ . This fact should not, however, be thought of as a disadvantage if the amplitude of fluctuation of the  $\log Y$  values around the  $\log \hat{Y}$  values is constant over time. By minimizing  $\sum (\log Y - \log \hat{Y})^2$ , cyclical deviations are allowed to exercise an influence in the trend more nearly equal to their relative magnitudes and tend to cause the trend to run approximately through the centers of the

cycles in early years, when the cyclical fluctuations may be of small amplitude in an absolute sense, but large relative to the trend.

**Second-degree Exponential.** A second-degree exponential has the equation form

$$\hat{Y} = AB^XC^{X^2}$$

or in logarithmic form and in terms of  $X$  values

$$\log \hat{Y} = a + bX + cX^2$$

where  $a = \log A$ ,  $b = \log B$ , and  $c = \log C$ . In terms of  $x$  values

$$\log \hat{Y} = a + bx + cx^2$$

In logarithm form the equation is appropriately called a logarithmic parabola. The fitting of a logarithmic parabola is similar to the fitting of the second-degree polynomial equation, except that the first step is to look up the logarithms of the  $Y$  values, and the last step is to look up the antilogs of the  $\log \hat{Y}$  values. Table 19.10 illustrates the computation, and from it we find the necessary quantities to evaluate

$$b = \frac{\sum x \log Y}{\sum x^2} = \frac{4.09162}{28} = 0.146129$$

Using Eqs. (19-17), in logarithmic form, we find that

$$\begin{aligned} na + c \sum x^2 &= \sum \log Y \\ a \sum x^2 + c \sum x^4 &= \sum x^2 \log Y \\ 7a + 28c &= 11.53293 \\ 28a + 196c &= 44.47170 \end{aligned}$$

so that

and solving in the usual way

$$\begin{aligned} a &= 1.72661 \\ c &= -0.019762 \end{aligned}$$

we obtain

$$\log \hat{Y} = 1.72661 + 0.146129x - 0.019762x^2$$

and taking the antilogarithm of both sides, we see that

$$\hat{Y} = 53.3(1.4)^x(0.9554)^{x^2}$$

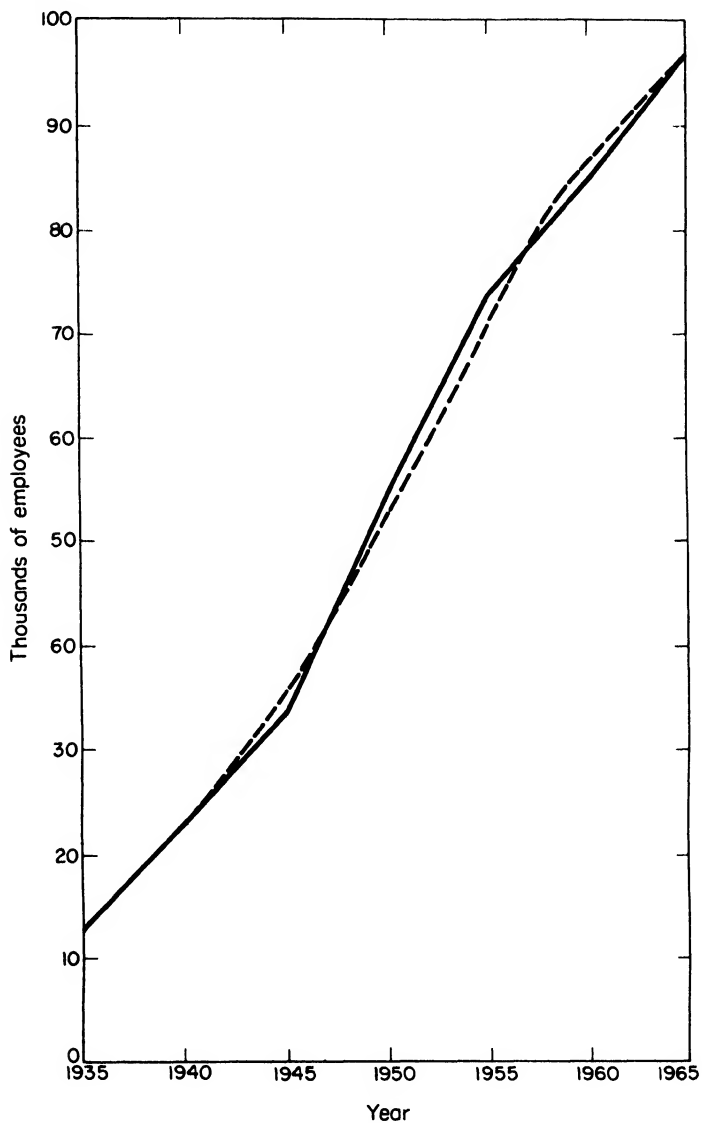
Typically,  $B$  and  $C$  are positive with  $B > 1$  and  $C < 1$ , as in this example. If we take the ratios of successive values of  $\hat{Y}$  (first ratios) and then successive ratios of these ratios (second ratios), we find these second ratios to have a constant value  $C^2$ . The quantity  $C^2 - 1$  is sometimes called the rate of retardation, because it is the percentage rate of decline in the first ratios. It

TABLE 19.10: CALCULATION OF SECOND-DEGREE EXPONENTIAL TREND FOR AVERAGE NUMBER OF PERSONS EMPLOYED IN THE RADIO BROADCASTING AND TELEVISION INDUSTRY, 1935-1965 (THOUSANDS OF EMPLOYEES)

Year	$X$	$x$	$x^2$	$Y$	$\log Y$	$x \log Y$	$x^2 \log Y$	$\log \bar{Y}$	Antilog ( $\log \bar{Y}$ ) trend value $\bar{Y}$
1935	0	-3	9	13	1.11394	-3.34182	10.02546	1.11036	13
1940	1	-2	4	23	1.36173	-2.72346	5.44692	1.35530	23
1945	2	-1	1	34	1.53148	-1.53148	1.53148	1.56072	36
1950	3	0	0	55	1.74036	0	0	1.72661	53
1955	4	1	1	74	1.86923	1.86923	1.86923	1.85298	71
1960	5	2	4	85	1.92942	3.85884	7.71768	1.93982	87
1965	6	3	9	97	1.98677	5.96031	17.88093	1.98714	97
Total	21	0	28	381	11.53293	4.09162	44.47170	11.53293	380
Mean	3	...	...	...	...	...	...	...	...
Sum of squares	...	...	196*	...	...	...	...	...	...

\*  $\Sigma x^2$ .Source: *Department of Commerce, The National Income and Product Accounts of the United States, 1929-1965, Statistical Tables, Washington, 1966, pp. 102ff.*

**CHART 19.8: AVERAGE NUMBER OF EMPLOYEES IN THE BROADCASTING INDUSTRY, 1935-1965, AND SECOND-DEGREE EXPONENTIAL TREND (ARITHMETIC SCALE).**



*Source: Table 19.10.*

is also true that the second differences of the  $\log \hat{Y}$  values are constant and equal to  $2c$ .

The data and trend for this example are plotted on arithmetic paper in Chart 19.8 and on logarithmic paper in Chart 19.9.

**Modified Exponential.** The modified exponential curve is an example of an asymptotic growth curve, as are the two remaining growth curves in this section. The modified exponential equation<sup>(8)</sup>

$$\hat{Y} = k + AB^X$$

has the property that the ratios of the first differences have the constant value  $B$ . The constant value  $k$  is an *asymptote*, or limit, which the trend values approach as  $X$  approaches  $\infty$  (or  $-\infty$ ). The asymptote may be either an upper or a lower limit, depending on the values of  $A$  and  $B$ .

The shape of the modified exponential curve depends upon the values of the constants  $k$ ,  $A$ , and  $B$ . Chart 19.10 shows the general shapes that the modified exponential can take when  $k$  is positive. Although it is mathematically possible for  $k$  to be negative, it is usually logically impossible for economic time series. The modified exponential can take any of the shapes shown on Chart 19.10. Typically  $A > 0$  and  $B > 1$ , so the diagram (1) in Chart 19.10 is most common.

The method of least squares is not often used for fitting a modified exponential, because it is too difficult in practice.

Almost always one of two methods is used: (1) the method of selected points, and (2) the method of partial totals. We shall illustrate these techniques using the data of Table 19.11.

TABLE 19.11: GENERAL EXPENDITURES OF STATE AND LOCAL GOVERNMENTS, 1950-1956, AND MODIFIED EXPONENTIAL FITTED BY METHOD OF SELECTED POINTS (BILLIONS OF DOLLARS)

Year	$X$	Expenditures $Y$	$B^X$	$AB^X$	$\hat{Y} = k + AB^X$
1950	0	12.3	1.000	3.20	12.30
1951	1	13.0	1.2051	3.86	12.96
1952	2	13.7	1.4522	4.65	13.75
1953	3	14.7	1.7500	5.60	14.70
1954	4	15.8	2.1089	6.75	15.85
1955	5	17.2	2.5414	8.13	17.23
1956	6	18.9	3.0626	9.80	18.90

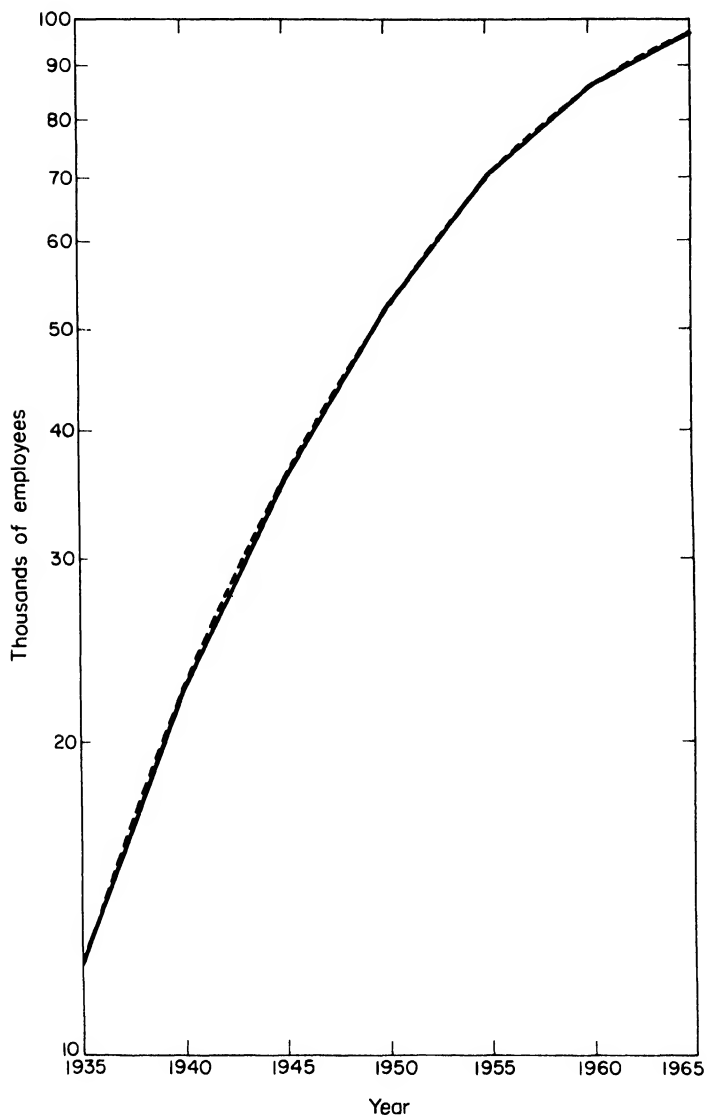
Source: U.S. Census Bureau, as reported in U.S. News and World Report, Vol. XLIII, July 12, 1957, p. 56.

<sup>(8)</sup> The modified exponential can also be put in linear form.

$$\log (\hat{Y} - k) = a + bX$$

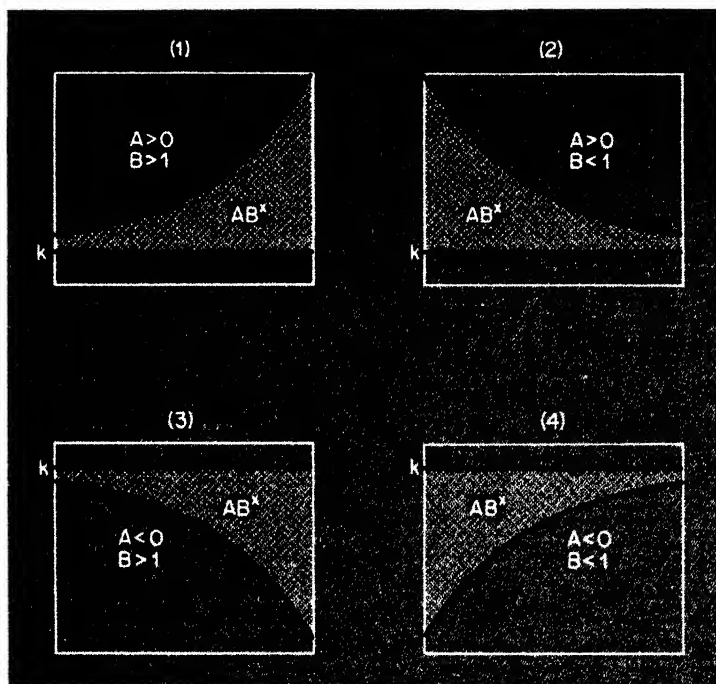
where  $a = \log A$  and  $b = \log B$ .

**CHART 19.9: AVERAGE NUMBER OF EMPLOYEES IN THE BROADCASTING INDUSTRY, 1935-1965, AND SECOND-DEGREE EXPONENTIAL TREND (LOGARITHMIC VERTICAL SCALE).**



*Source: Table 19.10.*

CHART 19.10: MODIFIED EXPONENTIAL CURVES.

**Method of Selected Points:**

(a) *Select the points.* Plot the data on paper with a logarithmic vertical scale, as in Chart 19.11, and draw a curve that gives a good fit. Read from the curve three values that are equidistant in time and call these values  $y_0$ ,  $y_1$ , and  $y_2$ . The  $y$  values need not be, and usually are not, the same as any of the observed  $Y$  values. The distance between  $x_0$  and  $x_1$ , and between  $x_1$  and  $x_2$  is  $r$  years, where  $r = (n - 1)/2$ . If possible, the years should be the first, middle, and last years. For our example  $n = 7$  and  $r = 3$ . The selected points are

$$1950: x_0 = 0; \quad y_0 = 12.3$$

$$1953: x_1 = 3; \quad y_1 = 14.7$$

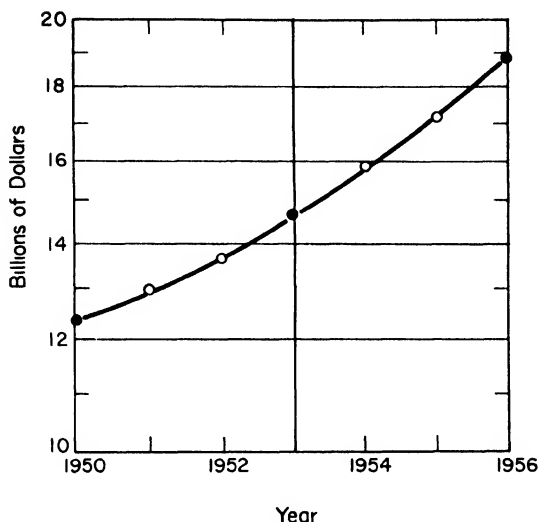
$$1956: x_2 = 6; \quad y_2 = 18.9$$

(b) *Compute the constants.* The equations are

$$Br = \frac{y_2 - y_1}{y_1 - y_0} \quad (19-25)$$

$$A = \frac{y_1 - y_0}{Br - 1} \quad (19-26)$$

$$k = y_0 - A \quad (19-27)$$

**CHART 19.11: GENERAL EXPENDITURES OF STATE AND LOCAL GOVERNMENTS, 1950-1956, AND FREE-HAND TREND.**

Source: Table 19.11.

For our illustration we have

$$r = \frac{7 - 1}{2} = 3$$

$$B^3 = \frac{18.9 - 14.7}{14.7 - 12.3} = \frac{4.2}{2.4} = 1.75$$

$$B = \sqrt[3]{B^3} = \sqrt[3]{1.75} = 1.20507$$

$$A = \frac{2.4}{1.75 - 1} = \frac{2.4}{0.75} = 3.2$$

$$k = 12.3 - 3.2 = 9.1$$

The trend equation is

$$\hat{Y} = 9.1 + 3.2(1.20507)^x$$

and selected  $\hat{Y}$  values are evaluated in Table 19.11.

**Method of Partial Totals.** This method is more objective than the method of selected points and, therefore, less flexible.

(a) *Calculate partial totals.* Divide the data into three parts, each of which has  $r$  observations, where  $r = n/3$ . The totals for the three parts are respectively  $\Sigma_1 Y$ ,  $\Sigma_2 Y$ ,  $\Sigma_3 Y$ . The criterion of fit is satisfied when  $\Sigma_1 Y = \Sigma_1 \hat{Y}$ ,  $\Sigma_2 Y = \Sigma_2 \hat{Y}$ , and  $\Sigma_3 Y = \Sigma_3 \hat{Y}$ . The partial totals are shown in Table 19.12.



**TABLE 19.12: GENERAL EXPENDITURES BY STATE AND LOCAL GOVERNMENTS, 1950-1955, AND MODIFIED EXPONENTIAL TREND FITTED BY METHOD OF PARTIAL TOTALS (BILLIONS OF DOLLARS)**

Year	Expenditures $Y$	$B^x$	$AB^x$	$\hat{Y} = k + AB^x$
1950	12.3	1.0000	2.8878	12.33
1951	13.0	1.2181	3.5176	12.96
$\Sigma_1 Y$	25.3	...	...	25.3
1952	13.7	1.4838	4.2849	13.73
1953	14.7	1.8074	5.2194	14.66
$\Sigma_2 Y$	28.4	...	...	28.4
1954	15.8	2.2016	6.3578	15.80
1955	17.2	2.6818	7.7445	17.19
$\Sigma_3 Y$	33.0	...	...	33.0

Source: See Table 19.11.

(b) *Compute the constants.* The equations are

$$B^r = \frac{\Sigma_3 Y - \Sigma_2 Y}{\Sigma_2 Y - \Sigma_1 Y} \quad (19-28)$$

$$A = (\Sigma_2 Y - \Sigma_1 Y) \frac{B - 1}{(B^r - 1)^2} \quad (19-29)$$

$$k = \frac{1}{r} \left[ \Sigma_1 Y - A \left( \frac{B^r - 1}{B - 1} \right) \right] \quad (19-30)$$

For the data in Table 19.12

$$r = \frac{9}{3} = 3$$

$$B^3 = \frac{33.0 - 28.4}{28.4 - 25.3} = 1.48387$$

$$B = 1.2181$$

$$A = (28.4 - 25.3) \frac{0.2181}{(0.48387)^2} = 2.8878$$

$$k = \frac{1}{3} \left[ 25.3 - 2.8878 \left( \frac{0.48387}{0.2181} \right) \right] = 9.4466$$

Notice that the method of partial totals as we have applied it is severely limited in that it presupposes a number of observations that is some multiple of three. The method of selected points, on the other hand, presupposes an odd number of observations.

**Gompertz.** The Gompertz curve, which is usually stated in the form

$$\hat{Y} = kGB^x$$

can also be stated in a form that is analogous to the modified exponential<sup>(9)</sup>

$$\log \hat{Y} = \log k + AB^X$$

where  $A = \log G$ .

The first differences of the  $\log \hat{Y}$  values have a constant ratio  $B$ . Typically, the value of  $A$  is negative, and the value of  $B$  is between zero and one. Therefore, the  $\log \hat{Y}$  values have the shape of diagram (4) of Chart 19.10. Hence, when plotted on paper with a logarithmic vertical scale the Gompertz curve is concave from below and thus indicates that the percentage growth is getting smaller. Typically also, a Gompertz curve has the shape of a non-symmetrical S curve when plotted on arithmetic paper. The lower asymptote is zero, and the upper asymptote is  $k$ . The curve is said to be nonsymmetrical because its behavior on opposite sides of the point of inflection is different. Because of its nonsymmetrical character, the first differences of Gompertz trend values form a curve resembling a positively skewed frequency distribution.

The Gompertz curve can be fitted by the method of selected points to obtain  $B^r$ ,  $A$ ,  $\log k$ , and  $\log \hat{Y}$  by use of formulas identical with Eqs. (19-25) through (19-27) except that  $\log y$  is substituted for  $y$  throughout. It may also be fitted by the method of partial totals by using formulas identical to Eqs. (19-28) through (19-30), except that  $\log Y$  is substituted for  $Y$  throughout.

**Logistic.** The logistic curve is usually stated in the form

$$\hat{Y} = \frac{k}{1 + 10^{a' + bX}} \quad (19-31)$$

where  $a' = \log(ka)$ . The equation can also be stated in a form that is analogous to the modified exponential. If we take the reciprocal of both sides of Eq. (19-31),

$$\frac{1}{\hat{Y}} = \frac{1}{k} + \frac{10^{a' + bX}}{k}$$

and let  $A = (1/k)10^{a'} = A'/k$  and  $B = 10^b$ , so that

$$\frac{1}{\hat{Y}} = \frac{1}{k} + AB^X \quad (19-32)$$

the analogy is clear.

For this latter equation, the first differences of the reciprocals of the trend have a constant ratio  $B$ . Typically, the value of  $A$  is positive and the value

<sup>(9)</sup> The Gompertz curve can be written in linear form.

$$\log \log \left( \frac{\hat{Y}}{k} \right) = a + bX$$

where  $a = \log A$  and  $b = \log B$ .

The logarithms may be taken to the base  $e$ , instead of the base 10, without changing the values of  $k$  or  $\hat{Y}$ . The values of  $a$  and  $b$  will, however, be changed.

of  $B$  is between zero and one. Therefore, the reciprocals of a logistic curve have the shape of diagram (2) of Chart 19.10. When plotted on paper with a logarithmic vertical scale, the logistic curve is concave from below and thus indicates that the percentage growth is getting smaller. Typically, also, a logistic curve has the shape of a symmetrical S curve. The lower asymptote is zero, the upper asymptote is  $k$ , and the behavior of the curve on both sides of its point of inflection is the same. The first differences resemble an approximately normal frequency distribution.

Until the point of inflection is reached, the amount of growth of a logistic curve gets larger as  $\hat{Y}$  gets larger, but eventually this is counteracted by another force: the amount of growth gets smaller as  $\hat{Y}$  approaches its upper limit. We are reminded of the Malthusian law that population would grow geometrically if it were not for the limitation imposed by the scarce means of subsistence. For this and other reasons the logistic curve is often used to describe population growth and series closely connected with the size of the population. Since it has been used extensively by the biometricians R. Pearl and L. J. Reed, it is sometimes called the Pearl-Reed curve.

The logistic curve may be fitted by the method of selected points to obtain  $B^r$ ,  $A$ ,  $1/k$ , and  $1/\hat{Y}$ . The formulas are identical to Eqs. (19-25) through (19-27), except that  $1/y$  is substituted for  $y$  throughout the calculation.<sup>(10)</sup> It may also be fitted by the method of partial totals using formulas identical to Eqs. (19-28) through (19-30), except that  $1/Y$  is substituted for  $Y$  throughout.

## 19.9 MOVING AVERAGE AS TREND

A centered moving average is a series of averages of  $r$  consecutive values, each average being placed at the chronological center of these consecutive values. Thus, for a five-year moving average of the data in

---

<sup>(10)</sup> The logistic equation can also be stated

$$\hat{Y} = \frac{k}{1 + e^{a' + bX}}$$

which changes the values of  $a'$  and  $b$  but not  $k$  or  $\hat{Y}$ . Also, the logistic can be stated in linear form in various ways. For example, from Eq. (19-32),

$$\log \left( \frac{1}{\hat{Y}} - \frac{1}{k} \right) = a + bX$$

also

$$\log \left( \frac{k - \hat{Y}}{\hat{Y}} \right) = a' + bX$$

In this second form we can show that the relative growth is a linear function of the level attained. This is perhaps the most interesting concept of the logistic curve.

Table 19.13 we compute

$$1942: \frac{1.3 + 2.2 + 0.8 + 0.3 + 0.5}{5} = 1.02$$

$$1943: \frac{2.2 + 0.8 + 0.3 + 0.5 + 1.4}{5} = 1.04$$

and so on.

**TABLE 19.13: INDUSTRIAL PURCHASES OF NEW STRUCTURES, 1940-1965, AND ILLUSTRATIONS OF MOVING AVERAGES (BILLIONS OF 1958 DOLLARS)**

<i>Year</i>	<i>Y</i> ( <i>New construction purchases</i> )	<i>5-Year moving average (centered)</i>	<i>4-Year moving average (uncentered)</i>	<i>4-Year moving average (centered)</i>
1940	1.3	...	...	...
1941	2.2	...	...	...
1942	0.8	1.02	1.15	1.05
1943	0.3	1.04	0.95	0.85
1944	0.5	1.20	0.75	1.02
1945	1.4	1.56	1.30	1.59
1946	3.0	1.88	1.88	.
1947	2.6	2.04	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
1959	2.1	2.70	.	.
1960	2.8	2.56	.	.
1961	2.7	2.62	2.60	2.68
1962	2.8	2.86	2.75	2.82
1963	2.7	3.20	2.88	3.10
1964	3.3	...	3.32	...
1965	4.5	...	...	...

*Source: Department of Commerce, The National Income and Product Accounts of the United States, 1929-1965, Statistical Tables, Washington, 1966, pp. 82-83.*

When the moving average is of an even number of items, the procedure is somewhat more laborious, since centering requires another step: a two-term moving average of the  $r$  term moving average. Thus, if we took a four-term moving average of the data of Table 19.13, the first four-term moving average would be centered between 1941 and 1942, and the second would be centered between 1942 and 1943. A two-term moving average of these two four-term moving averages would be centered at 1942. Thus

$$1941-1942: \frac{1.3 + 2.2 + 0.8 + 0.3}{4} = 1.15$$

$$1942-1943: \frac{2.2 + 0.8 + 0.3 + 0.5}{4} = 0.95$$

and  $1942: \frac{1.15 + 0.95}{2} = 1.05$

and so on. Obviously, we could accomplish the same result by taking an  $r + 1$  term moving average, if  $r$  is even, giving the end values in each moving

average half the weight of the other values, and dividing by the sum of the weights<sup>11</sup>

$$1942: \frac{1(1.3) + 2(2.2) + 2(0.8) + 2(0.3) + 1(0.5)}{8} = 1.05$$

Occasionally a time series will have a trend that cannot be fitted adequately by a simple mathematical curve. In such a case a moving average may be used. Although a moving average becomes smoother as  $r$  is increased when the fluctuations around the trend are random, this statement must be modified if the fluctuations are periodic. In this case  $r$  should coincide with an integral multiple of the length of the periodic movements. Consider the hypothetical data below. The symbol  $\bar{Y}_4$  is used to mean centered four-year moving average. The residual fluctuations apparently have a period of four years about the linear trend.

We have seen that if the cycles are uniformly of length  $r$ , an  $r$ -term moving average will completely smooth out the cycles, leaving only the trend, if the trend is linear. But since cycles in economic time series are not usually of uniform length, the number of observations should approximate the average

$Y$	$\bar{Y}_4$	$Y - \bar{Y}_4$
2	...	...
3	...	...
8	6	2
9	7	2
6	8	-2
7	9	-2
12	10	2
13	11	2
10	12	-2
11	13	-2
16	...	...
17	...	...

<sup>(11)</sup> In general, if  $\bar{Y}_r$  represents the centered  $r$ -term moving average of a set of  $Y_i$  values

$$\bar{Y}_r = \sum_{j=1}^k W_j Y_{i+j-1}$$

for  $i = 1, 2, \dots, n - k + 1$ . When  $r$  is odd  $k = r$ . When  $r$  is even,  $k = r + 1$ . When  $r$  is odd,  $W_j = 1/r$  for all  $W_j$ . When  $r$  is even,  $W_1 = W_{r+1} = 1/2r$  with the remainder of the  $W_j$ -values equal to  $1/r$ . Thus, the first 5-term moving average in Table 19.13 can be calculated

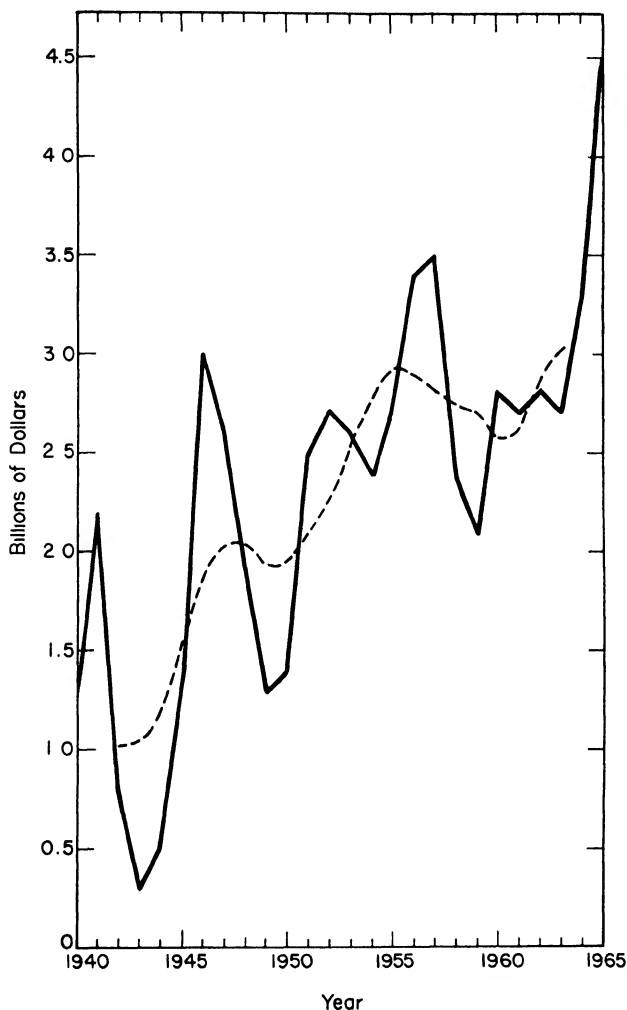
$$1942: 0.20(1.3) + 0.20(2.2) + 0.20(0.8) + 0.20(0.3) + 0.20(0.5) = 1.02$$

since  $1/r = \frac{1}{5} = 0.20$  and  $r$  is odd. The first centered 4-term moving average in Table 19.13 can be calculated

$$1942: 0.125(1.3) + 0.25(2.2) + 0.25(0.8) + 0.25(0.3) + 0.125(0.5) = 1.05$$

since  $1/r = \frac{1}{4} = 0.25$  and  $1/2r = \frac{1}{8} = 0.125$ .

**CHART 19.12: INDUSTRIAL PURCHASES OF NEW STRUCTURES, 1940-1965, AND 5-YEAR MOVING AVERAGE (BILLIONS OF 1958 DOLLARS).**



*Source: Table 19.13.*

length of one cycle, or perhaps two consecutive cycles. Actually, this rule is hard to follow in practice, because the lengths of cycles vary greatly in duration. Virtually all that we can do is proceed by trial and error until a trend that is visually satisfactory is obtained.

Chart 19.12 shows the original data and the five-year moving average of Table 19.13. The time series seems to show cyclical fluctuations with period

of about five years. Notice, however, that the five-term moving average is of an undulatory character. Presumably a ten-year moving average would straighten the trend somewhat.

Moving averages have several inherent defects that prevent their common use.

1. They never cover the complete period. If  $r$  is odd,  $(r - 1)/2$  trend values are lost at each end of the data. If  $r$  is even,  $r/2$  values are lost at each end of the data.

2. Moving averages cannot easily be described by mathematical equations. They cannot be thought of as "laws" of growth.

3. Simple moving averages tend to follow the data into those peaks and troughs that are of the greatest amplitude and duration.

4. Frequently, moving averages misrepresent the level of the trend. For instance, if the trend is an exponential curve, the moving average will smooth out part of the curvature.

## 19.10 SELECTION OF TREND TYPE AND PERIOD TO WHICH TO FIT TREND

In the selection of a curve to express the trend of a time series, a wide range of choice is possible, and it is not always easy to decide which type of curve is most appropriate. Although not completely satisfactory, the following considerations are useful.

1. In general, a trend that can be expressed as an equation with time as the independent variable is to be preferred to a moving average. Even though it cannot be said that the series conforms to an economic "law," such an equation is useful for purposes of description, summarization, and comparison.

2. Before any computations are undertaken, the data should be plotted on arithmetic paper. If the points fall approximately on a straight line, a straight line equation is indicated. If there is one bend, a second-degree equation is indicated. If there is a point of inflection, a third-degree equation is indicated.

3. Also plot the data on semilogarithmic paper. If the points fall approximately on a straight line, an exponential curve is indicated. If there is one bend, a second-degree exponential is indicated.

4. If successive differences or successive ratios of the  $Y$  values, the  $\log Y$  values or the  $1/Y$  values are approximately constant, the trend type is indicated. In order to determine whether an exponential, a second-degree exponential, a modified exponential, a Gompertz, or a logistic is appropriate, the following preliminary test is suggested.

First, plot the data on semilogarithmic paper and draw a freehand curve that gives a good fit. Read from the curve at least five values that are

equidistant in time. The first, middle, and last points are those used in fitting a modified exponential, Gompertz, or logistic.

If the selected values fall on a straight line when plotted on semilogarithmic paper, or if the ratios of successive values fall on a horizontal straight line when plotted on arithmetic paper, an exponential is indicated.

If the ratios of successive values fall on a straight line when plotted on semilogarithmic paper, or if the second ratios fall on a horizontal straight line when plotted on arithmetic paper, a second-degree exponential is indicated.

If the ratios of the first differences (but not the ratios of successive values) fall on a horizontal straight line when plotted on arithmetic paper, a modified exponential is indicated.

If the ratios of the first differences of the logarithms fall on a horizontal straight line when plotted on arithmetic paper, a Gompertz curve is indicated.

If the ratios of the first differences of the reciprocals fall on a horizontal straight line when plotted on arithmetic paper, a logistic curve is indicated.

5. If the  $(Y - k)$  values,  $(\log Y - \log k)$  values, or  $(1/Y - 1/k)$  values are approximately a straight line when plotted on semilogarithmic paper, there is a strong presumption in favor of the modified exponential, the Gompertz, or the logistic, respectively. Experimentation with arbitrary values of  $k$ ,  $\log k$ , and  $1/k$  may reveal that the data are well described by one of the above curves over part, but not all, of the time covered.

6. If a logical reason can be assigned for a series behaving in a manner described by a particular equation, preference should be given to that trend, even though it is not intended to project the trend. If the trend is to be projected ahead a short period (one or two years), it is particularly important that a trend be selected that will behave in a logical fashion. Thus, if the business is thought to be reaching a saturation point and will, when that point is reached, remain stationary, a trend should be selected that will flatten out at the top—not one that will continue upward or soon bend downward. If, on the other hand, it is believed that the rate of increase that has persisted will be continued, an exponential curve should be chosen.

7. In order to determine the general shape of the curve desired, the statistician must understand the underlying economic factors. He must be able to distinguish trends from the various other movements present in the series. Having decided which variations are which, he is in a position to select a type of trend that will conform to these distinctions. If he can make up his mind which trend shows the cycles most accurately, he will have one criterion for deciding which trend to select.

8. Other things being equal, select a trend equation with as few constants as possible, bearing in mind that a trend is not satisfactory unless points 6 and 7 are satisfied. If there are too many constants in the equation selected, too many degrees of freedom will be used up, and the degrees of freedom remaining for the unexplained variation will be so few that some of the



constants will not be significant. (In the extreme case, where there are as many constants as observations, the trend will go through each observation, but it will be meaningless.) One should also remember that the different observations in an economic time series are interdependent, and therefore there are never as many *independent* observations (degrees of freedom) as there are items to begin with.

The period of time to which the trend is fitted often makes considerable difference. If the data cover only 10 or 15 years it may be important to consider the stage of the cycle at the initial and terminal years; for longer periods this is less important. Consider the data of Table 19.14. The true trend equation is

$$\hat{Y} = 8.5 + 1x$$

with origin between year 4 and year 5. This equation can be obtained by fitting a trend to the data of years 1 through 8 or 3 through 6. The cyclical residuals are of a periodic nature, the length of each cycle being 4 years. The data and three trend lines are plotted on Chart 19.13. The solid line is the true trend.

Using these data as an example, we can make certain generalizations.

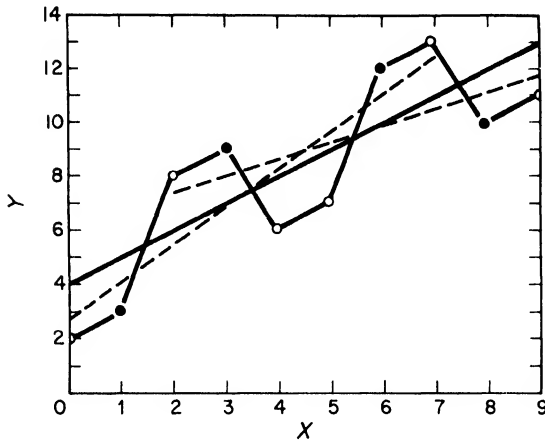
1. The first year should not be one of depression and the last year one of prosperity, for that will make  $b$  too large. Thus, fitting to years 0–7, we obtain  $b = 1.38$ . This trend, plotted on Chart 19.13, is too steep. Conversely, the first year should not be one of prosperity and the last year one of depression, for that will make  $b$  too small. Thus, fitting to years 2–9, we obtain  $b = 0.62$ . This trend is also plotted on Chart 19.13 and is not steep enough. (Note that the average value of the two  $b$  values is 1, which is the correct  $b$  value.)

2. The beginning and ending years should not usually be at the same point in their respective cycles. Thus if we fit the trend to years 0–8, which are

TABLE 19.14: ARTIFICIAL DATA, TREND, AND CYCLICAL RESIDUALS

Year $X$	$X - \bar{X}$ $x$	Data $Y$	Trend $\hat{Y}$	Cyclical residuals $Y - \hat{Y}$
0	-4.5	2	4	-2
1	-3.5	3	5	-2
2	-2.5	8	6	2
3	-1.5	9	7	2
4	-0.5	6	8	-2
5	0.5	7	9	-2
6	1.5	12	10	2
7	2.5	13	11	2
8	3.5	10	12	-2
9	4.5	11	13	-2

CHART 19.13: ARTIFICIAL DATA AND THREE TRENDS.



first years of depressions,  $b = 1.13$ . If we fit to years 1–9, which are second years of depression,  $b = 0.87$ . These discrepancies are not so spectacular as those of the preceding paragraph, but they are substantial.

3. If the number of depression years exceeds the number of prosperity years, the level of the trend will usually tend to be too small even though the slope may be correct. Thus, if we fit to years 0–9, the trend equation is

$$\hat{Y} = 8.1 + 1x$$

Thus  $a = 8.1$  instead of 8.5. If the number of prosperity years exceeds the number of depression years, the level of the trend usually will be too large, even though the slope may be correct. Thus, if we fit to years 2–7, the trend equation is

$$\hat{Y} = 9.2 + 1x$$

Thus  $a$  is 9.2 instead of 8.5.

4. If the number of prosperity years is the same as the number of depression years, and if the first and last years are on opposite sides of the cycle (not opposite sides of the trend), both the level  $a$  and the slope  $b$  will tend to be correct. As stated earlier, if the trend is fitted to years 1–8 or 3–6, the correct trend will be obtained.

In general, bias in the slope is more to be feared than bias in the level, for a trend becomes cumulatively worse as  $|x|$  increases if  $b$  is wrong. None of the trends we have mentioned have a value of 8.5 when  $X = 4.5$  (and  $x = 0$ ) except those fitted to years 1–8 or 3–6. However, it is only for trends fitted to years 0–9 or 2–6 that the discrepancy is noteworthy.

## PROBLEMS

1. Fit a first- and second-degree polynomial to the following series. Which trend seems more satisfactory? What can you tell by inspection of first and higher differences of the  $Y$  values?

$x$	-9	-7	-5	-3	-1	1	3	5	7	9
$Y$	265.1	283.1	286.8	339.4	407.6	407.5	457.1	435.7	586.9	604.6

2. Fit a growth curve that you consider to be a satisfactory representation of the trend in the following series.

$x$	-9	-7	-5	-3	-1	1	3	5	7	9
$Y$	265.1	283.1	286.8	339.4	407.6	407.5	457.1	435.7	586.9	604.6

3. Show that the equation in footnote 11 can be used to take first differences.

4. Explain the following statements.

*a. Choice of the term of a moving average involves a tradeoff between smoothness and flexibility.*

*b. A centered  $r$  term moving average is the linear least squares regression estimate of the central term of the data included in the given moving average.*

## Time Series Analysis: Seasonal, Cyclical, and Irregular Movements

In the previous chapter we discussed one component of the time series model given by Eq. (19-1), the secular trend  $T$ . In this chapter we will discuss the remaining components of that model, the seasonal  $S$ , cyclical  $C$ , and irregular  $I$  movements.

| Seasonal variation is a type of recurring fluctuation that has a duration of one year. Changes in the weather—amount of daylight, temperature, humidity rainfall, wind velocity—affect consuming habits and producing ability. Certain holidays, notably Christmas and Easter, likewise are important factors in business fluctuations occurring within a year.

✓ There are various reasons for measuring seasonal movements. If it is known that the price of some commodity typically fluctuates in a particular fashion, it may be profitable to buy while the price is low, holding the commodity for subsequent use or sale. Before deciding on such a course one must know the cost of storage and other costs involved. Similarly, if it is desirable to have stable production within a year even though sales show seasonal fluctuation, one can plan his fluctuating inventory so as to have at all times a large enough stock on hand to take care of expected sales, and yet keep his inventory close to the minimum. If, on the other hand, it seems better to keep small inventories, allowing production of a particular commodity to fluctuate, it may still be possible to maintain a

constant labor force by manufacturing commodities the sales of which are complementary in their seasonal movements. Finally, it may be possible to reduce the seasonal fluctuations in sales by a proper advertising or price policy, or in production by artificial control of temperature, humidity, and so on, in the plant.

A less obvious reason for measuring seasonal movements is to adjust the data statistically for such movements, thus leaving the series composed only of trend, cyclical movements, and irregular fluctuations. Data in this form are easier to interpret for many purposes, since one is less likely to confuse the reason for any observed movement. For instance, if the data have not been seasonally adjusted, one may mistake a seasonal upswing for an improvement in business conditions; or one may become pessimistic when the reason for a decline is merely the usual seasonal slump. When data have been properly adjusted for seasonal movements, there should be no tendency for any one month of a year to be higher or lower than the average of the two adjacent months, or for any pattern to repeat itself in successive years. The method of making this adjustment, if the model is multiplicative, is to divide the original data by the seasonal index. The data are then said to be *deseasonalized*.

$$\text{Deseasonalized data: } \frac{Y}{S} = \frac{TCSI}{S} = TCI$$

Two types of seasonal movement are (1) the stable seasonal, the pattern of which is constant for a period of years, and (2) the changing seasonal, the pattern of which is gradually changing during a period of time. Sometimes a stable pattern is maintained for a number of years, after which the pattern suddenly changes. In that case two stable seasonal indexes should be computed.

Another type of fluctuation that is considered in the chapter is that which is attributable to calendar variation. This variation might be considered a quasi-seasonal type of variability.

Business cycles are undulatory movements representing alternating rises and declines in business phenomena. For example, there seem to be cycles in interest rates, prices, inventory holdings, employment, etc. Business cycles are not strictly repetitious movements, and hence are perhaps more appropriately called business fluctuations.

Irregular movements are classifiable into two types: (1) random and (2) specific. Random irregular movements are the result of a large number of small independent causes. Individual random deviations are not predictable, but considered as a whole they may reasonably be expected to have an approximately normal distribution with mean of zero. Their occurrence over time also should be random; the sign of any deviation should not have any bearing on the sign of the next deviation. In statistical jargon, we say that there should not be any significant autocorrelation.

Specific irregular movements are movements for which we can assign a specific cause. For example, a sharp drop in steel production might be assignable to a strike of steel workers, a drop in stock prices to an illness of the president of the United States, and so on.

Cyclical movements differ from random fluctuations in that cyclical observations for successive months are interdependent. Runs above the trend line, runs below the trend line, runs up, and runs down are not only longer than can be accounted for by chance but also longer than for specific irregular movements. In statistical jargon we say that most economic time series, even after adjustment for trend and seasonal, have significant autocorrelation.

## 20.1 CALENDAR VARIATION

**Adjustment for Length-of-month Variation.** When a time series, reported on a monthly basis, represents the sum of daily values it is often a good idea to adjust the monthly series for variability in length of months. If we wish to adjust for length-of-month variability, without disturbing the *level* of the monthly data, we may multiply the original data by a length-of-month adjustment factor, which is

$$\frac{\text{Average number of days per month}}{\text{Actual number of days per month}}$$

For example, in an ordinary year there are 365 days, so the average number of days per month is

$$\frac{365}{12} = 30.41667$$

Therefore, for any January of an ordinary year, the adjustment factor is

$$\frac{30.41667}{31} = 0.981183$$

since January has 31 days. The procedure is the same for a leap year, except that there are 366 days in all. If one adjustment factor is made for each month, for each year of a time series, the series which results from multiplying each of the monthly observations of the original series by the length-of-month adjustment factors is said to be adjusted for length-of-month variation. Often no adjustment is made for length-of-month variation; it is considered part of seasonal variation.

**Adjustment for Trading-day Variation.** There is even greater variation in the number of trading days than there is in calendar days. There may be either four or five Sundays and four or five Saturdays in a month. Needless to say, a variation in production or sales brought about by an increase or decrease in working days is of limited economic significance. If

the data are to be put on a trading-day basis, one must first obtain the number of trading days per month by subtracting the number of Saturdays and/or Sundays from the calendar days in each month. Also, Saturdays and Sundays may count as full holidays, half-holidays, or no holiday at all, depending upon the industry or trade in question.

The number of trading days for each month of each year having been obtained, production or sales per trading day are obtained by dividing the production or sales figures by the number of trading days. Once the number of trading days in each month of each year has been computed, one multiplies the original data for each month by the trading-day adjustment factor.

$$\frac{\text{Average number of trading days per month}}{\text{Actual number of trading days per month}}$$

Adjustment for length-of-month and/or trading-day variation is facilitated by use of Appendix 16, where one will find a flexible arrangement that permits one quickly to ascertain the number of calendar or trading days in any month from 1898 to 1976.

Modern computer-programmed seasonal-adjustment procedures usually provide options whereby one can choose whether to: (1) make no adjustment for calendar variation; (2) adjust only for length-of-month variation; (3) adjust for trading-day variation.<sup>(1)</sup> Calendar variation adjustment may be done in one operation, or in two stages: (1) adjustment for length-of-month variation; (2) adjustment for trading-day variation not accounted for by length-of-month variation. The procedure is to regress the estimated irregular series upon the number of times each day of the week occurs in each particular month. From the seven resulting weights (regression coefficients) monthly factors are constructed and divided into the data to remove the trading-day variation. In order to obtain the estimated irregular series it is necessary to obtain a preliminary estimate of  $TC$  and a preliminary estimate of  $S$ .

## 20.2 A CONVENTIONAL METHOD FOR ESTIMATING A STABLE SEASONAL COMPONENT

If a time series does not contain a trend or cyclical component and has a stable, unchanging, seasonal component, the seasonal component may be estimated by calculating the average value of all Januaries, all Februaries, etc. These averages are usually expressed as percentages, so that the twelve percentages will average 100 percent. The reason we want the seasonal index to average 100 percent is that we may wish to deseasonalize the original data

---

<sup>(1)</sup> See, for example, U.S. Department of Commerce Bureau of the Census, Technical Paper No. 15, *The X-11 Variant of the Census Method II Seasonal Adjustment Program*, especially pp. 2-3. For a computer program see: Lawrence Salzman, *Computerized Economic Analysis* (New York: McGraw-Hill Publishing Company, Inc., 1968).

(by dividing by the seasonal index). If the seasonal index averages 100 percent, the deseasonalized data will have approximately the same average value as the original data.<sup>(2)</sup>

Since few, if any, economic time series conform to the assumptions set forth in the above paragraph, the method of simple averages of monthly values is seldom used. Instead, an attempt to eliminate the trend and the cyclical components of the time series is made before averaging the different months. Briefly, the following method is used:

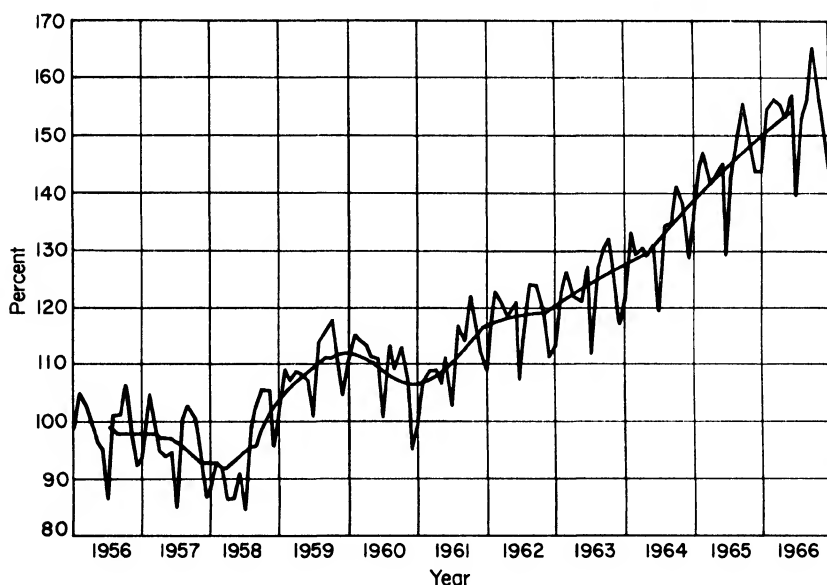
1. Estimate  $TC$ , using a centered 12-month moving average.
2. Eliminate  $TC$ , obtaining estimates of  $SI$

$$SI = \frac{TCSI}{TC}$$

3. Average the  $SI$  estimates by months, obtaining an estimate of  $S$ .

Table 20.1 and Chart 20.1 show the Federal Reserve Index of Home Goods and Apparel, by months, 1956 through 1966. In Table 20.1 the middle

**CHART 20.1: FEDERAL RESERVE INDEX OF HOME GOODS AND APPAREL AND CENTERED TWELVE MONTH MOVING AVERAGE, 1956-1966.**



Source: Table 20.1.

<sup>(2)</sup> This same procedure can be carried out by using "dummy" variables and multiple regression analysis. See Michael C. Lovell, "Seasonal Adjustment of Economic Time series and Multiple Regression Analysis," *Journal of the American Statistical Association*, Vol. 58, 1963, pp. 993-1010. This article is also informative on the problem of loss of degrees of freedom resulting from seasonal adjustment.

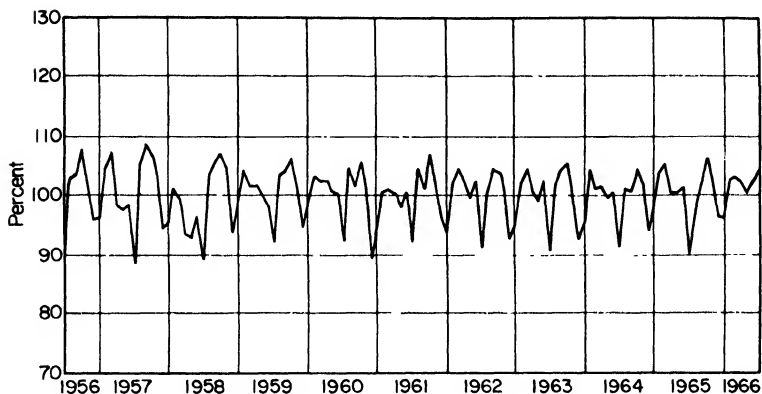


**TABLE 20.1: FEDERAL RESERVE INDEX OF HOME GOODS AND APPAREL, 1956-1966, AND COMPUTATION OF PERCENTAGES OF CENTERED 12-MONTH MOVING AVERAGE**

<i>Year and month</i>	<i>Index 1957-1959 = 100 TCSI</i>	<i>13-month weighted sum</i>	<i>12-month centered moving average</i>	<i>Percent of moving average</i>
(1)	(2)	(3)	(4)	(5)
1956: Jan.	99.8	...	...	...
Feb.	105.4	...	...	...
March	103.1	...	...	...
April	100.5	...	...	...
May	96.4	...	...	...
June	95.6	...	...	...
July	86.7	2371.1	98.80	87.8
Aug.	101.3	2362.2	98.43	102.9
Sept.	101.9	2360.7	98.36	103.6
Oct.	106.0	2357.5	98.23	107.9
Nov.	97.8	2350.3	97.93	99.9
Dec.	93.9	2347.7	97.82	96.0
1957: Jan.	94.1	2346.0	97.75	96.3
Feb.	102.2	2343.9	97.66	104.6
March	104.8	2343.9	97.66	107.3
April	95.6	2339.2	97.47	98.1
May	94.1	2333.1	97.21	96.8
.	.	.	.	.
.	.	.	.	.
1965: Aug.	142.1	3481.7	145.07	98.0
Sept.	148.9	3500.4	145.85	102.1
Oct.	155.8	3522.7	146.78	106.1
Nov.	149.8	3546.9	147.79	101.4
Dec.	144.0	3568.8	148.70	96.8
1966: Jan.	144.0	3589.0	149.54	96.3
Feb.	154.5	3608.7	150.36	102.8
March	156.0	3626.7	151.11	103.2
April	155.3	3643.4	151.81	102.3
May	153.3	3659.6	152.48	100.5
June	156.4	3668.9	152.87	102.3
July	138.9	...	...	...
Aug.	152.8	...	...	...
Sept.	156.2	...	...	...
Oct.	165.2	...	...	...
Nov.	156.6	...	...	...
Dec.	146.5	...	...	...

*Col. (2): Source, Board of Governors of the Federal Reserve System, Federal Reserve Bulletin, various issues. Col. (3): Sum of each consecutive 13 values with 11 central values counted twice. Col. (4): Col. (3) values divided by 24. Col. (5): 100 [Col. (2)/Col. (4)].*

**CHART 20.2: PERCENTAGES OF TWELVE-MONTH MOVING AVERAGE OF FEDERAL RESERVE INDEX OF HOME GOODS AND APPAREL.**



*Source: Table 20.1.*

part of the data is omitted for purposes of condensation. The line running through Chart 20.1 is a centered 12-month moving average which is a preliminary estimate of the *TC* movements. It was shown in Chapter 19 that if the number of observations in each average of a moving average is the same as the length of the periodic movements we wish to smooth out, the desired smoothing will be realized. Therefore, if seasonal swings are periodic and last 12 months, a 12-month moving average will entirely eliminate these seasonal movements. At the same time the moving average will largely smooth out irregular movements. Unfortunately, it will also smooth out a small amount of the cyclical movements. It is mainly for that reason that we speak of a 12-month moving average as being only an *estimate* of the *TC* movements. Computations are given in Table 20.1.

In column (5) of Table 20.1 are shown the percentages of the centered 12-month moving average. These percentages are plotted on Chart 20.2 and represent the *SI* estimates. Because the 12-month moving average smooths out a little of the cycle at peaks and troughs, these estimates of *SI* are slightly biased, having a slight positive correlation with the cyclical movements.

Before we decide whether to compute a single seasonal index, two or more seasonal indexes for different periods of time, or a moving seasonal (one that gradually changes pattern), it is well to study Chart 20.2. A casual inspection of that chart does not seem to indicate any great change in seasonal pattern. Therefore, we shall compute a single seasonal index for these data.

The estimates of *SI* shown in Table 20.2 are taken directly from column (5) of Table 20.1. If the irregular component of the *SI* values is a random

TABLE 20.2: FEDERAL RESERVE INDEX OF HOME GOODS AND APPAREL, ESTIMATE OF SEASONAL IRREGULAR MOVEMENTS AND COMPUTATION OF STABLE SEASONAL INDEX BY CONVENTIONAL METHOD (PERCENTAGE OF 12-MONTH MOVING AVERAGE)

Year	Jan.	Feb.	Mar.	April	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.	Total
1956	...	...	...	...	...	...	87.8L	102.9	103.6	107.9H	99.9L	96.0	...
1957	96.3	104.6H	107.3H	98.1	96.8	98.3	88.5	105.0H	108.1H	107.0	104.1	94.5	...
1958	95.4	101.0L	99.6L	93.3L	92.9L	96.7L	89.2	103.0	105.2	107.0	104.7H	93.7	...
1959	98.5	104.1	101.4	101.5	100.2	98.6	92.2	103.5	104.3	106.2	100.8	94.5	...
1960	98.9H	103.1	102.4	102.3H	100.9H	100.3	92.3	104.7	101.4	105.3	100.9	89.4L	...
1961	93.2L	100.7	101.0	100.5	98.0	100.3	92.5H	104.4	101.0	107.0	102.4	96.6	...
1962	93.8	102.1	104.7	102.3	99.6	102.3	91.1	100.9	104.3	104.0L	100.7	92.9	...
1963	94.6	102.1	104.3	100.3	99.1	102.7H	90.7	101.7	104.3	105.3	101.1	92.9	...
1964	95.7	104.1	101.2	101.3	99.8	100.4	91.1	101.1	100.8L	104.7	101.5	94.1	...
1965	97.7	104.0	105.2	100.5	100.3	101.3	90.0	98.0L	102.1	106.1	101.4	96.8H	...
1966	96.3	102.8	103.2	102.3	100.5	102.3	...	...	...	...	...	...	...
Selected sum	768.3	823.0	823.4	806.8	794.3	803.8	725.1	822.2	826.2	848.6	812.9	755.2	9609.8
Modified mean	96.04	102.88	102.92	100.85	99.29	100.48	90.64	102.78	103.28	106.08	101.61	94.40	1201.22
Seasonal index	95.94	102.78	102.82	100.75	99.19	100.38	90.55	102.68	103.17	105.97	101.51	94.30	1200

Source: Table 20.1, Col. (5).

Selected sum: Values labeled H (High) and L (Low) have been discarded.

Modified mean: Selected sum divided by 8.

Seasonal index: Modified mean multiplied by correction factor:  $0.99898 = 1200/1201.22$ .

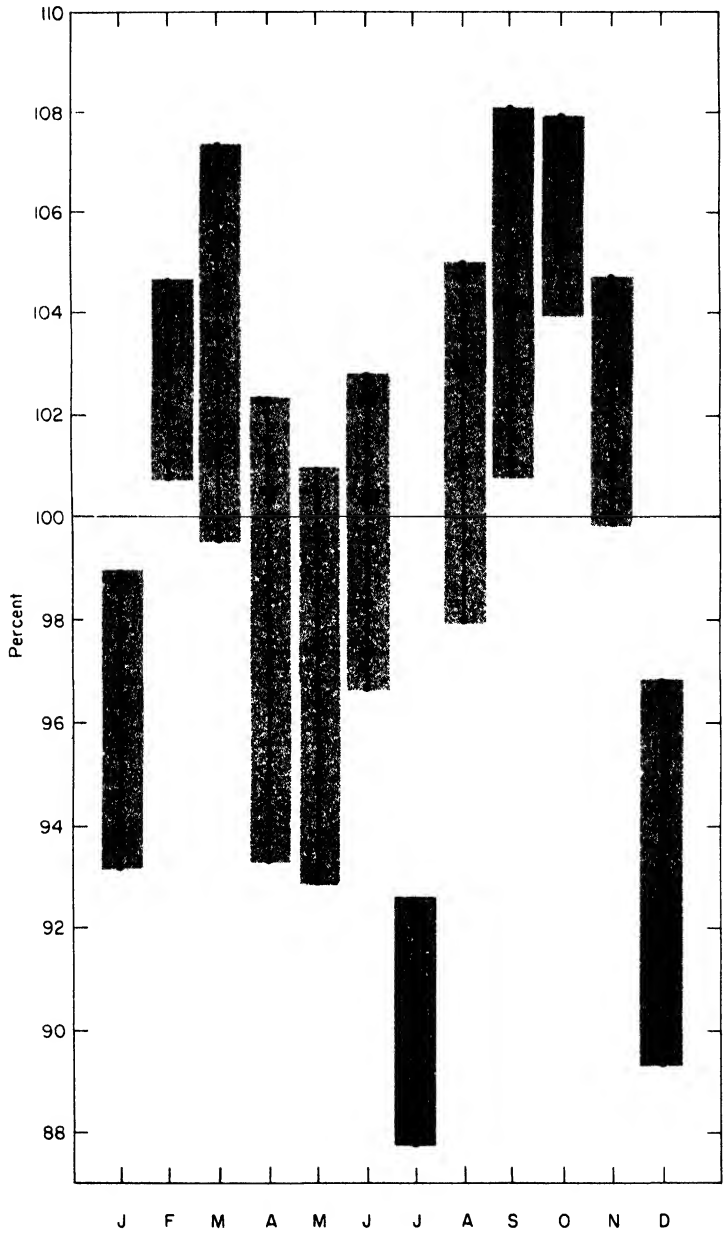
variable, normally distributed, with mean of zero, then monthly means of the *SI* values will average out the irregular movements, leaving only the seasonal component. In practice the elimination of the irregular movements by averaging is done in various ways. Some statisticians compute a weighted arithmetic mean for each month, in which the weights are proportional to the "credibility" of the *SI* estimates. Credibility is taken to be a function of  $z$ , the number of standard deviations a value deviates from the simple arithmetic mean of the monthly arrays. Other statisticians completely disregard values which are "too extreme," or for which an assignable cause can be found, and compute the simple arithmetic mean of the remaining items. This procedure is really a variation of the first method, the weights being either one or zero. Perhaps the most common of all simple methods of eliminating the irregular component in the *SI* estimates is the one we shall illustrate. A subjectively determined number of extreme values is discarded from each array, usually the highest and lowest, or the two highest and two lowest, and the arithmetic mean of the remaining values is computed. Such a mean is called a modified mean. The extreme case of a modified mean would be one in which all except one or two observations are discarded. Here, the resulting statistic is the median. The justification for these methods is that *SI* is not a random variable but that it contains nonrandom irregularities and that *SI* values near cyclical highs or cyclical lows are biased because of the bias in the 12-month centered moving average.

Chart 20.3 shows the *SI* estimates of Table 20.1, column (5). After inspection of this chart, where it is seen that for several months there is a single value which departs from the bulk of the other values, it was decided to eliminate the largest and smallest *SI* value in each monthly array and to compute modified means from the eight central values of each month. This procedure has been followed in Table 20.2. Here the largest and smallest *SI* values in each month have been discarded and a modified mean computed from the eight central values for each month. The sum of these modified means is 1201.22. This figure is very close to 1200 (or an average of 100.0). When we multiply each modified mean by the correction factor 0.99898, we obtain seasonal index numbers that total 1200.0, and average 100.0.

Table 20.3 shows the original *Y* values and the stable seasonal index of Table 20.2. The seasonally adjusted *Y* values are shown in Table 20.4. The seasonal adjustment, again, is accomplished by dividing each *Y* value by the appropriate seasonal index number. It is somewhat easier to multiply each of the *Y* values by a *seasonal adjustment* factor, which is the reciprocal of the seasonal index number.

Because a 12-month moving average tends to smooth out part of the cyclical movements, it is sometimes considered worthwhile to make a revised estimate of *TC*. This may be done by computing a five-month moving average

**CHART 20.3: SEASONAL-IRREGULAR MOVEMENTS OF FEDERAL RESERVE INDEX OF HOME GOODS AND APPAREL, 1954-1966.**



*Source: Table 20.2.*

TABLE 20.3: FEDERAL RESERVE INDEX OF HOME GOODS AND APPAREL, STABLE SEASONAL INDEX, AND STABLE SEASONAL ADJUSTMENT FACTORS

Year	Jan.	Feb.	Mar.	April	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.	Total
1956	99.8	105.4	103.1	100.5	96.4	95.6	86.7	101.3	101.9	106.0	97.8	93.9	1,188.4
1957	94.1	102.2	104.8	95.6	94.1	95.3	85.3	100.6	102.6	100.6	97.1	87.7	1,160.0
1958	88.3	93.4	92.1	86.4	86.6	90.8	84.6	99.0	102.5	105.9	105.5	95.9	1,131.0
1959	102.2	109.4	107.7	108.8	108.2	107.2	100.9	113.9	115.3	117.9	112.3	105.5	1,309.3
1960	110.5	115.2	114.2	113.6	111.6	110.4	100.8	113.5	109.4	113.1	107.9	95.4	1,315.6
1961	99.5	107.8	108.4	108.5	106.5	110.1	102.6	116.7	114.0	121.9	117.7	111.9	1,325.6
1962	109.2	119.3	112.9	120.5	117.6	120.8	107.7	119.6	124.0	123.9	120.1	111.2	1,416.8
1963	113.6	123.1	126.3	122.1	121.2	126.2	112.0	126.3	130.1	131.9	127.3	117.5	1,477.6
1964	121.6	132.9	129.8	130.5	129.5	131.2	120.0	134.2	135.1	141.6	138.3	129.2	1,573.9
1965	135.1	144.6	147.2	141.8	142.6	145.2	129.9	142.1	148.9	155.8	149.8	144.0	1,727.0
1966	144.0	154.5	156.0	155.3	153.3	156.4	138.9	152.8	156.2	165.2	156.6	146.5	1,835.7
Total	1217.9	1307.8	1312.5	1283.6	1267.6	1289.2	1169.4	1320.0	1340.0	1383.8	1330.4	1238.7	15,460.9
Seasonal index	95.94	102.78	102.82	100.75	99.19	100.38	90.55	102.68	103.17	105.97	101.51	94.30	1,200
Seasonal adjustment factor	1.04232	0.97295	0.97257	0.99256	1.00817	0.99621	1.10436	0.97390	0.96927	0.94366	0.98512	1.06045	...

Source: Table 20.1.

Stable seasonal index: Table 20.2.

Seasonal Adjustment factors:  $1/(\text{Stable seasonal index numbers})$ .

TABLE 20.4: FEDERAL RESERVE INDEX OF HOME GOODS AND APPAREL ADJUSTED BY STABLE SEASONAL INDEX

Year	Jan.	Feb.	Mar.	April	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.	Total
1956	104.0	102.5	100.3	99.8	97.2	95.2	95.7	98.7	98.8	100.0	96.3	99.6	1,188.1
1957	98.1	99.4	101.9	94.9	94.9	94.9	94.2	98.0	99.4	94.9	95.7	93.0	1,159.3
1958	92.0	90.9	89.6	85.8	87.3	90.5	93.4	96.4	99.4	99.9	103.9	101.7	1,130.8
1959	106.5	106.4	104.7	108.0	109.1	106.8	111.4	110.9	111.8	111.3	110.6	111.9	1,309.4
1960	115.2	112.1	111.1	112.8	112.5	110.0	111.3	110.5	106.0	106.7	106.3	101.2	1,315.7
1961	103.7	104.9	105.4	107.7	107.4	109.7	113.3	113.7	110.5	115.0	115.9	118.7	1,325.9
1962	113.8	116.1	119.5	119.6	118.6	120.3	118.9	116.5	120.2	116.9	118.3	117.9	1,416.6
1963	118.4	119.8	122.8	121.2	122.2	125.7	123.7	123.0	126.1	124.5	125.4	124.6	1,477.4
1964	126.7	129.3	126.2	129.5	130.6	130.7	132.5	130.7	130.9	133.6	136.2	137.0	1,573.9
1965	140.8	140.7	143.2	140.7	143.8	144.6	143.5	138.4	144.3	147.0	147.6	152.7	1,727.3
1966	150.1	150.3	151.7	154.1	154.6	155.8	153.4	148.8	151.4	155.9	154.3	155.4	1,835.8
Total	1269.3	1272.4	1276.4	1274.1	1278.2	1284.2	1291.3	1285.6	1298.8	1305.7	1310.5	1313.7	15,460.2

Source: Entries in Table 20.3 multiplied by seasonal adjustment factors. A check on computation is provided by the totals.

of the deseasonalized data.<sup>(3)</sup> After this is done the original data are divided by the five-month moving average. These percentages of five-month moving average are considered the revised *SI*, and a seasonal index is computed from the revised estimates, the method of Table 20.2 being used.

### 20.3 A SIMPLE METHOD FOR ESTIMATING A MOVING SEASONAL COMPONENT

There can be no reasonable doubt concerning the existence of seasonal fluctuations in the home goods and apparel series. Whether the stable seasonal index we have calculated is adequate is another matter. There are several tests of the adequacy of seasonal adjustment, and we will discuss three of them.<sup>(4)</sup> First, it would seem reasonable to require that a seasonally adjusted series should remain unchanged by a second seasonal adjustment. Second, and somewhat easier to calculate, is a test which forms the ratio of each month to the average of the preceding and following months after the data have been deseasonalized. For example, the ratio for September, 1956 would be found by forming the percentage

$$100 \left[ \frac{98.9}{(98.7 + 100.0)/2} \right] = 99$$

The September ratios for all years are given as follows:

1956	1957	1958	1959	1960	1961	1962	1963	1964	1965	1966
99	103	101	101	98	97	103	102	99	101	99

These ratios should average 100 for each month (as they do for September), and there should be no "significant runs" above or below 100 for any month. In September the only "significant" run seems to be three items in a row above 100. Third, a rather recent test involves spectral analysis of the

<sup>(3)</sup> The simplest satisfactory moving average is a 5-month moving average. Both the U.S. Census Bureau and the U.S. Bureau of Labor Statistics use more complicated moving averages sometimes involving as many as 23 terms. The Federal Reserve System has used a technique which is somewhat more subjective. The technique is explained and illustrated in: H. C. Barton, Jr., "Adjustment for Seasonal Variation," *Federal Reserve Bulletin*, June, 1941.

<sup>(4)</sup> The first test has been called the "idempotency" test by Michael C. Lovell, *op. cit.* Lovell discusses other tests as well. For the ratio of current month to the average of the preceding and following months see: Henry A. Latané, "Seasonal Factors Determined by Difference from Average of Adjacent Months," *Journal of the American Statistical Association*, Vol. 37, 1942, pp. 517-522. For the spectral analysis approach see: Marc Nerlove, "Spectral Analysis of Seasonal Adjustment Procedures," *Econometrica*, Vol. 32, 1964, pp. 241-286.



TABLE 20.3: ESTIMATE OF MOVING SEASONAL INDEX FOR FEDERAL RESERVE INDEX OF HOME GOODS AND APPAREL

Year	Jan.	Feb.	Mar.	April	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
1956	96.9	102.6	101.6	100.0	98.3	99.1	90.0	103.7	104.4	106.7	101.9	94.2
1957	96.9	102.6	101.6	100.0	98.3	99.1	90.0	103.7	104.4	106.7	101.9	94.2
1958	96.9	102.6	101.6	100.0	98.3	99.1	90.0	103.7	104.4	106.7	101.9	94.2
1959	96.9	102.6	101.6	100.0	98.3	99.1	91.2	104.2	103.6	106.7	102.5	94.2
1960	95.9	102.0	101.6	101.4	99.3	99.7	91.9	103.6	103.3	106.2	101.4	93.7
1961	95.6	102.4	102.7	101.4	99.6	101.0	91.9	103.2	103.3	105.6	100.9	93.4
1962	94.7	102.4	102.6	101.4	99.5	101.0	91.5	102.4	102.2	105.1	101.2	93.3
1963	94.7	102.7	103.4	100.8	99.5	101.3	91.0	101.2	102.5	105.4	101.3	94.5
1964	95.5	103.0	104.1	101.4	99.9	102.0	91.0	101.2	102.5	105.4	101.3	94.5
1965	95.5	103.0	104.1	101.4	99.9	102.0	91.0	101.2	102.5	105.4	101.3	94.5
1966	95.5	103.0	104.1	101.4	99.9	102.0	91.0	101.2	102.5	105.4	101.3	94.5

Source: Table 20.2.

variance (power) content of a time series before and after seasonal adjustment. Successful seasonal adjustment is that which removes power at "seasonal frequencies" without altering the spectrum elsewhere. This latter technique is mathematically difficult and, like the others, involves a good deal of subjective judgment. In the present case there is some indication that the stable seasonal index is inadequate, and we shall proceed to calculate a moving seasonal index.

The usual method for estimating a moving-seasonal component is to use *SI* estimates like those of Table 20.2 and to fit trends, by months, to these data. Thus, there will be twelve sets of trend values. There are various methods of fitting these trends.

1. A polynomial trend of appropriate degree may be fitted to each monthly array, the same degree of equation usually being used for each month. This procedure has the advantage that a seasonal estimate is provided for each month of each year. Also, if the *SI* estimates are adjusted to total 1200 for each year prior to fitting the trends, the *S* estimates will also total 1200 (average 100) each year. A possible disadvantage of this method is that the polynomial equations may not accurately describe the seasonal movements.

2. A moving average may be fitted to the *SI* estimates for each monthly array. A simple five-year moving average provides a reasonably satisfactory estimate of the *S* values. A somewhat more complicated five-year moving average is one that has for weights: 1, 2, 3, 2, 1. The sum of the weights is 9, and the moving average is usually referred to as a three of a three (a three-year moving average of a three-year moving average). The Census Bureau makes use of this type of moving average.

3. If polynomial trends or moving averages are fitted, it may be desirable to take account of extreme values. Thus values thought to be extreme may be given reduced weight or even assigned zero weight. This procedure is cumbersome and is not used in this text.

The method that we will illustrate here is one that uses a five-year modified moving average. The method uses the central three of each consecutive five years, eliminating the largest *SI* value and the smallest *SI* value for each consecutive five years of a given month. For example, in Table 20.5 January, 1959 was found, from Table 20.2, by evaluating

$$\frac{96.3 + 95.4 + 98.5}{3} = 96.9$$

Notice that 98.9 and 93.2 were not included in the average because they are the two extreme observations. January, 1960 was found by evaluating

$$\frac{95.4 + 98.5 + 93.8}{3} = 95.9$$

and so on for other months. Although a five-year modified moving average does not provide us with values for the first two years and the last two years,

the first two years are arbitrarily given the value of the third year, and the last two years are given the value of the year preceding. This procedure has a conservative bias, since it does not assume a continuation of the trends indicated by the moving average. A less conservative bias will result if one artificially extends the *SI* values before fitting the moving average. Thus, we might use the averages of the first two actual *SI* values in any array for the backward extension. The mean of the last two actual *SI* values in any array may be used for the forward extension.

4. Trends may be fitted visually. This method not only provides *S* values for each month of each year, but it allows the statistician to weight each *SI* value in accord with its credibility. But of course such a method is highly subjective.

**TABLE 20.6: FEDERAL RESERVE INDEX OF HOME GOODS AND APPAREL DESEASONALIZED BY MOVING SEASONAL INDEX**

<i>Year and month</i>	<i>Index 1957-1959 = 100</i>	<i>Estimate of moving seasonal index</i>	<i>Deseasonalized data</i>
(1)	(2)	(3)	(4)
1956: Jan.	99.8	96.9	103.0
Feb.	105.4	102.6	102.7
Mar.	103.1	101.6	101.5
April	100.5	100.0	100.5
May	96.4	98.3	98.1
June	95.6	99.1	96.5
.	.	.	.
.	.	.	.
.	.	.	.
1966: July	138.9	91.0	152.6
Aug.	152.8	101.2	151.0
Sept.	156.2	102.5	152.4
Oct.	165.2	105.4	156.7
Nov.	156.6	101.3	154.6
Dec.	146.5	94.5	155.0

*Column (2): From Table 20.1.*

*Column (3): From Table 20.5.*

*Column (4):  $100 [\text{Col. (2)} / \text{Col. (3)}]$ .*

Table 20.6 shows the adjustment of the original data by the estimated moving seasonal index. There is a separate seasonal index number for each month of each year.

## 20.4 MORE COMPLICATED METHODS OF ESTIMATING A MOVING SEASONAL COMPONENT

The method used to estimate the moving seasonal component, which was given in the last section, is quite simple as compared with techniques

currently in use in various governmental agencies. The Bureau of the Census and the Bureau of Labor Statistics have, over the years, progressively refined techniques of seasonal adjustment that are suitable for electronic computer calculation. Among the newest of these techniques are (1) the *X-11* variant of the Census II Seasonal Adjustment Program and (2) the BLS (Bureau of Labor Statistics) Seasonal Factor Method (1966). Bulletins of the above titles are published by the United States Department of Commerce and the United States Department of Labor, respectively.

## 20.5 ESTIMATING CYCLICAL MOVEMENTS

Cyclical movements are generally estimated by the following technique:

1. Estimate *TCI* by dividing the original data by the seasonal index numbers for each month.

$$TCI = \frac{Y}{S}$$

2. Estimate *CI* by dividing the *TCI* estimates by the trend estimates for each month.

$$CI = \frac{TCI}{T}$$

In Chart 20.4 the original data, as given in Chart 20.1, are plotted, and an estimate of the trend is represented by the dotted line. The original data do not seem to conform to any of the trend types in the last chapter. The trend was estimated by fitting one linear regression line to the monthly values of the years 1956–1961 and a second linear regression line for the monthly values of the years 1961–1966. The regression lines intersect at January, 1962.<sup>(5)</sup>

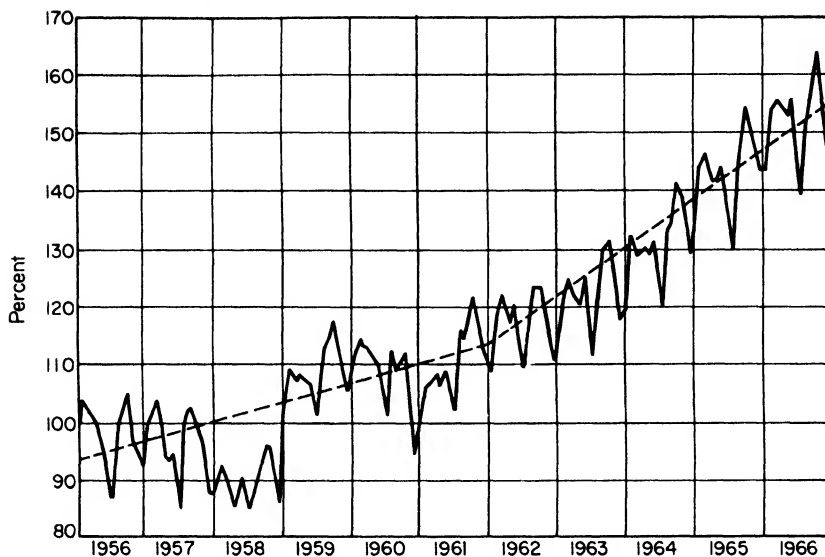
The trend values of Chart 20.4 are given in Table 20.7 as well as the estimated moving seasonal index numbers for each month. In columns (5) and (6) the estimates of *TCI* and *CI* are given, respectively.

In order to make estimates of the cyclical movements we must remove the irregular movements from the *CI* movements. We cannot do this by dividing *CI* values by *I* because the irregular movements are what remain after adjusting our original data for *T*, *C*, and *S*.

---

<sup>(5)</sup> Sometimes the trend is fitted to the yearly averages of the monthly values. With a small number of observations this method sometimes avoids bias in slope of the trend resulting from the seasonal movement.

**CHART 20.4: FEDERAL RESERVE INDEX OF HOME GOODS AND APPAREL AND ESTIMATE OF TREND, 1956-1966.**



First Trend Line:  $\hat{Y} = 93.8 + 0.264578X$

Second Trend Line:  $\hat{Y} = 104.8 + 0.7085281X$

Source: Table 20.1.

**TABLE 20.7: FEDERAL RESERVE INDEX OF HOME GOODS AND APPAREL, AND COMPUTATION OF CI MOVEMENTS USING TS ESTIMATES, 1956-1966**

Year and month	Index TCSI	Trend T	Seasonal S	TCI	CI
(1)	(2)	(3)	(4)	(5)	(6)
1956: Jan.	99.8	93.8	96.9	103.0	109.8
Feb.	105.4	94.1	102.6	102.7	109.2
Mar.	103.1	94.3	101.6	101.5	107.6
April	100.5	94.6	100.0	100.5	106.2
May	96.4	94.9	98.3	98.1	103.3
June	95.6	95.1	99.1	96.5	101.4
July	86.7	95.4	90.0	96.3	101.0
Aug.	101.3	95.7	103.7	97.7	102.1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
1966: Sept.	156.2	153.0	102.5	152.4	99.6
Oct.	165.2	153.7	105.4	156.7	102.0
Nov.	156.6	154.4	101.3	154.6	100.1
Dec.	146.5	155.1	94.5	155.0	100.0

Col. (2): from Table 20.1.

Col. (3): see Chart 20.4.

Col. (4): from Table 20.5.

Col. (5):  $100.0 [\text{Col. (2)} / \text{Col. (4)}]$ .

Col. (6):  $100 [\text{Col. (5)} / \text{Col. (3)}]$ .

To estimate  $C$ , the usual procedure is to take a moving average of  $CI$  in the hope that the average will smooth out the irregular movements. The method is not very satisfactory, but it is perhaps the best we can do. In the choice of the moving average two not completely independent decisions must be made: (1) the length of the moving average and (2) the system of weights to use in computing the moving average.

**Length of Moving Average.** Often a three-month or five-month moving average is used, because they are easy to compute and because they smooth out only a negligible part of the cycle. On the other hand, moving averages of short length are not very smooth. If a longer simple moving average is used, the cyclical estimates will be smoother, but the tendency to smooth out part of the cycle will be considerable.

**System of Weighting.** Three systems of weighting will be discussed:<sup>(6)</sup> (1) simple, (2) binomial, (3) polynomial. Simple moving averages, of odd length, assign the same weight to each item in the average. All previous moving averages illustrated in this text have been simple.

Binomially weighted moving averages employ binomial coefficients as weights. For a five-term binomially weighted moving average the weights are

$$1, 4, 6, 4, 1$$

and the sum of these weights is 16. Notice that the central item in the average receives the greatest weight and the weights assigned to the other items get gradually smaller as the item's position departs from the central position. The result is that any unusually large (or small) item has a small influence on the average when it is first included in the average, gradually increases in influence, and then gradually fades away.

Polynomial weights, as the name implies, are polynomials. The weights are used in an attempt to preserve the amplitude of the cycles and at the same time get smooth results. The weights are equivalent to fitting a polynomial trend of second- or higher-degree to each successive  $r$  items, and using the trend value at the central point of the moving average as the cyclical estimate. Such cyclical estimates are usually referred to as moving arcs.<sup>(7)</sup> Given below are some weights for five- and seven-term moving arcs.

---

<sup>(6)</sup> There are other systems of weighting also. For example, the U.S. Bureau of the Census frequently uses a 13-term Henderson with weights for the central month as follows:  $-0.019, -0.028, 0.000, 0.066, 0.147, 0.214, 0.240, 0.214, 0.147, 0.066, 0.000, -0.028, -0.019$ . If such a moving average is fitted to a parabola, the averages will fall exactly on the parabola. See U.S. Department of Commerce Bureau of the Census, Technical Paper No. 15, *The X-11 Variant of the Census Method II Seasonal Adjustment Program*, pp. 20 and 63.

<sup>(7)</sup> For a more complete explanation and list of weights see: Dudley J. Cowden, *Polynomial Moving Average Weights for Smoothing Irregular Fluctuations*, Technical Paper 5, School of Business Administration, University of North Carolina, 1965.

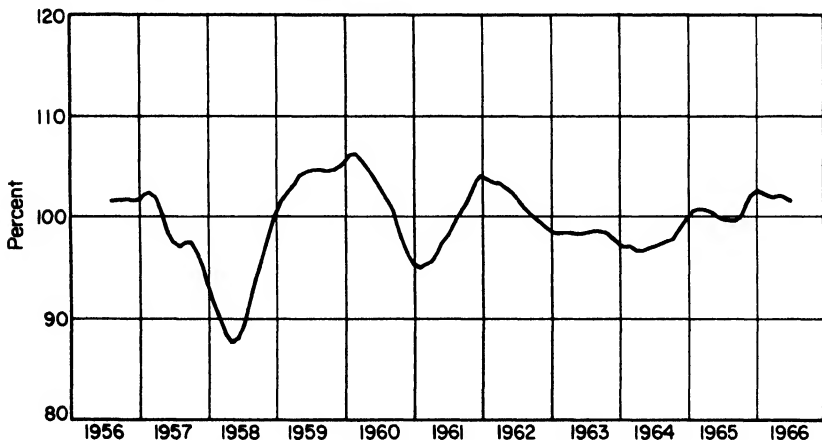
<i>5-term second degree</i>	<i>7-term second degree</i>	<i>7-term fourth degree</i>
—3	—2	5
12	3	—30
17	6	75
12	7	131
—3	6	75
	3	—30
	—2	5
<hr/> Total 35	<hr/> Total 21	<hr/> Total 231

Like binomial weights, moving arcs are relatively smooth. The function of the negative weights near the end of the weight pattern of a moving arc is to make the amplitude of the cyclical swings larger: to make them reach more faithfully into the cyclical peaks and troughs.

The five- and seven-year moving arcs are sufficiently sensitive, but on account of their short length, they are not always sufficiently smooth. If, however, two successive smoothings by a seven-month second-degree moving arc are used (i.e., if we take a seven-month moving arc of the first seven-month moving arc) the results are both sufficiently sensitive and sufficiently smooth. Such smoothing is illustrated in Table 20.8. For example, the first  $C_1$  value in that table is obtained as follows:

$$\frac{(-2)109.8 + (3)109.2 + (6)107.6 + (7)106.2 + (6)103.3 + (3)101.4 + (-2)101.0}{21}$$

**CHART 20.5: FEDERAL RESERVE INDEX OF HOME GOODS AND APPAREL CYCLICAL MOVEMENTS, 1956-1966.**



Source: Table 20.8.

and the first  $C_2$  value is obtained as follows:

$$\frac{(-2)105.7 + (3)103.5 + (6)102.0 + (7)101.2 + (6)101.9 + (3)101.8 + (-2)102.1}{21}$$

Other values of  $C_1$  and  $C_2$  are obtained in a similar manner. The  $C_2$  values are the estimates of the cyclical movements. These values are plotted on Chart 20.5.

## 20.6 IRREGULAR MOVEMENTS

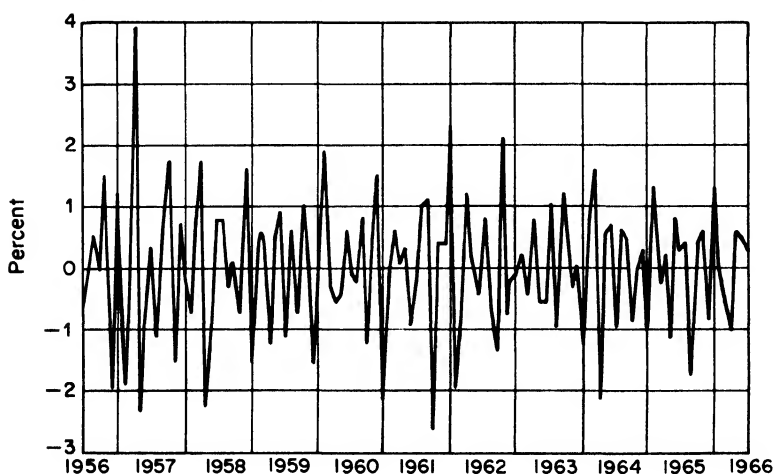
An estimate of the irregular movements may be obtained by dividing the cyclical-irregular estimates by the cyclical estimates

$$I = \frac{CI}{C}$$

and is done in Table 20.8, column (5). The last column in Table 20.8 shows the irregular movements as percentage deviations, and these deviations are plotted in Chart 20.6.

To the unaided eye the irregular movements in Chart 20.6 do not seem to form a regular pattern unless it is one of gradual decrease in the amplitude of the fluctuations over time. Also, there may be some evidence that the movements change direction more often than would be expected by chance. This latter condition might have arisen because too sensitive a moving

**CHART 20.6: FEDERAL RESERVE INDEX OF HOME GOODS AND APPAREL IRREGULAR MOVEMENTS, 1956-1966.**



Source: Table 20.8, Col. (6).



**TABLE 20.8: FEDERAL RESERVE INDEX OF HOME GOODS AND APPAREL CYCLICAL-IRREGULAR MOVEMENTS, CYCLICAL ESTIMATES, AND IRREGULAR MOVEMENTS, 1956-1966**

Year and month	CI	7-MONTH 2ND DEGREE MOVING ARC		$I$ (CI/C <sub>4</sub> ) percent	Percentage deviations (I - 100)
		C <sub>1</sub>	C <sub>2</sub>		
(1)	(2)	(3)	(4)	(5)	(6)
1956: Jan.	109.8	...	...	...	...
Feb.	109.2	...	...	...	...
Mar.	107.6	...	...	...	...
April	106.2	105.7	...	...	...
May	103.3	103.5	...	...	...
June	101.4	102.0	...	...	...
July	101.0	101.2	101.5	99.5	-0.5
Aug.	102.1	101.9	101.6	100.5	0.5
Sept.	101.8	101.8	101.8	100.0	0.0
Oct.	103.3	102.1	101.7	101.5	1.5
Nov.	99.6	101.5	101.5	98.1	-1.9
Dec.	103.1	101.0	101.8	101.2	1.2
1957: Jan.	100.1	102.3	102.1	98.1	-1.9
Feb.	102.5	102.9	102.3	100.1	0.1
Mar.	105.8	101.6	101.9	103.9	3.9
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
1965: July	99.8	99.3	99.4	100.4	0.4
Aug.	97.6	99.3	99.4	98.3	-1.7
Sept.	100.5	99.8	100.1	100.4	0.4
Oct.	101.8	101.2	101.2	100.6	0.6
Nov.	101.4	102.6	102.1	99.2	-0.8
Dec.	103.9	102.7	102.6	101.3	1.3
1966: Jan.	102.4	102.2	102.4	100.0	0.0
Feb.	101.4	101.9	101.9	99.5	-0.5
Mar.	100.8	101.5	101.8	99.0	-1.0
April	102.5	101.8	101.9	100.6	0.6
May	102.2	102.1	101.7	100.5	0.5
June	101.6	101.6	101.3	100.3	0.3
July	100.7	100.3	...	...	...
Aug.	99.1	100.1	...	...	...
Sept.	99.6	100.2	...	...	...
Oct.	102.0	...	...	...	...
Nov.	100.1	...	...	...	...
Dec.	100.0	...	...	...	...

Col. (2) from Table 20.7.

average was used to obtain the cyclical estimates, and there is, therefore, an indication that a moving average of greater length and/or different weights could profitably be tried to obtain revised cyclical estimates.

Irregular movements are sometimes tested for "randomness." One might, for example, form a frequency distribution of the irregular movements and utilize the chi square distribution to examine the goodness of fit of a

normal, log-normal, or some other probability distribution to the observed frequency distribution of the irregular movements. Many analysts test for autocorrelation prior to the test for goodness of fit to some probability distribution. Autocorrelation is explained in the next chapter.

---

## PROBLEMS

1. *a. Why is a centered 12-month moving average of the  $Y$  values considered to be only an estimate of the TC movements?*  
*b. What are some considerations in deciding whether to compute a stable seasonal index or two stable seasonal indexes or a moving seasonal index?*  
*c. Why can we not calculate  $C$  by dividing  $CI$  by  $I$ ?*  
*d. What are some practical reasons for seasonally adjusting an economic time series?*
2. How might you modify the methods of Sec. 20.2 if you were asked to seasonally adjust quarterly data?
3. Write down the weights for a seven-term binomially weighted moving average.
4. Develop a set of weights for obtaining  $C_2$  (as in Table 20.8) in one operation rather than two.

## Correlation of Time Series and Forecasting

All businessmen are forced to make forecasts, and decisions are constantly being made in the light of anticipated economic conditions. All forecasting techniques are designed either to limit the reliance upon judgment or to make judgment more reliable by providing significant facts and relationships. Although complete reliance cannot be placed upon any procedure, statistical or otherwise, any clue that may be helpful is worth noting.

The methods of forecasting discussed in this chapter are:

1. Forecasting a series by itself, sometimes referred to as the economic rhythm method.
2. Forecasting a series by other series, sometimes referred to as the cyclical sequence method.
3. Specific historical analogy.
4. Surveys of plans and opinions.
5. Crosscut economic analysis.
6. Simultaneous equation models.

### 21.1 CORRELATION OF CYCLICAL RELATIVES

Before time series are correlated it is usually desirable to eliminate trend and seasonal variations from both series, since

interest ordinarily centers on comparisons of cyclical changes.<sup>(1)</sup> If the trend is not removed, the correlation coefficient will indicate partly whether or not the trends of the two series are similar.<sup>(2)</sup> Similarity of trends, however, can better be determined merely by comparing the trends directly, either mathematically or graphically.

If we are interested in long-run relationships, a plausible case can be made in favor of correlating the data without adjustment for trend. It should be remembered, though, that similarity of trends is an extremely weak indication of causal relationship.

Failure to adjust the data for trend (if it is assumed that there are no seasonal movements or that seasonal movements have been eliminated) may make the correlation coefficient larger or smaller and may even change its sign. Consider the two artificial time series plotted as Chart 21.1a. Their cycles are similar, but their trends are in opposite directions. If we adjust each series for trend (by subtracting its trend values), we get the cycles shown by Chart 21.1b. Both series have exactly the same cycles, and the correlation between them is  $r = +1.0$ . If, however, we plot the *unadjusted* series as a scatter diagram, we get the results of Chart 21.1c. The correlation coefficient is  $r = -0.135$ . If we had correlated the unadjusted data for the years 2 through 5, the cycles would have been of more importance compared with the trend, and the correlation would have been positive. If, on the other hand, the series had been over a longer period of time, with the same tendencies in operation, the trend would have assumed even greater importance compared with the cycles, and the correlation coefficient would have had a still larger negative value.

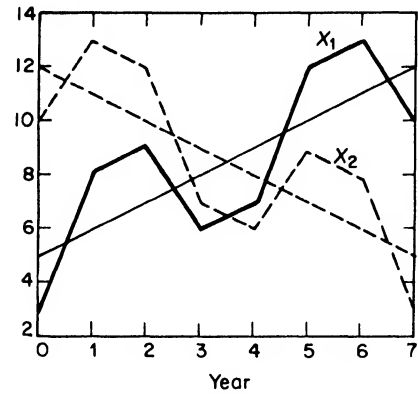
Consider also the two artificial time series plotted as Chart 21.2a. Their trends are the same, but their cycles are contrasting. If we adjust each series for trend (by subtraction), we get the cycles shown by Chart 21.2b. Their cycles fluctuate in opposite directions at the same time, and the correlation between them is  $r = -1.0$ . If, however, we plot the *unadjusted* series as a scatter diagram, we get the results of Chart 21.2c. The correlation coefficient is  $r = +0.135$ . If we had correlated the unadjusted data for the years 2 through 5, the cycles would have been of more importance compared with the trend, and the correlation would have been negative. If, on the other hand,

---

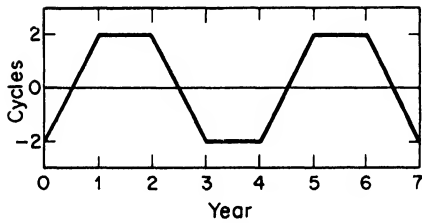
<sup>(1)</sup> Instead of removing trend the practice is sometimes followed of using time as a second independent variable in a multiple regression equation. The observations representing time  $x_s$  are usually consecutive numbers with 0 assigned to the central year or month. The numerical value of the *partial* correlation coefficient  $r_{12.3}$  is the same that would be obtained if each of the two series were adjusted by *subtracting* the straight line trend values and the *simple*  $r$  computed for the two *adjusted* series.

<sup>(2)</sup> If a pronounced seasonal is not removed from monthly data (and if it is assumed that the trends are unimportant), the coefficient will indicate to a considerable extent the similarity or dissimilarity of the two seasonal indexes instead of the relationship between the cyclical movements.

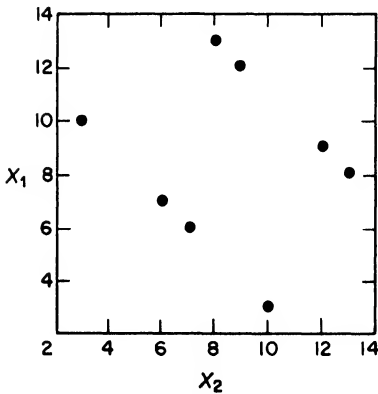
**CHART 21.1: TWO ARTIFICIAL TIME SERIES WITH IDENTICAL CYCLES, BUT WITH TRENDS HAVING OPPOSITE SIGNS.**



(a) Original data

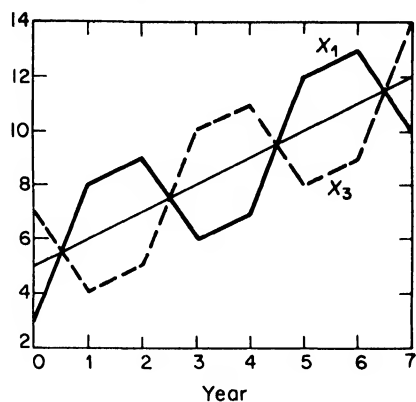


(b) Cycles

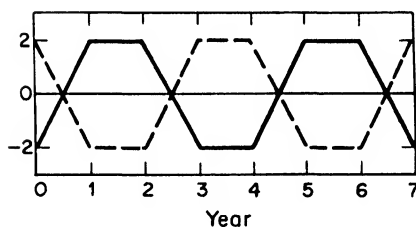


(c) Scatter diagram of original data

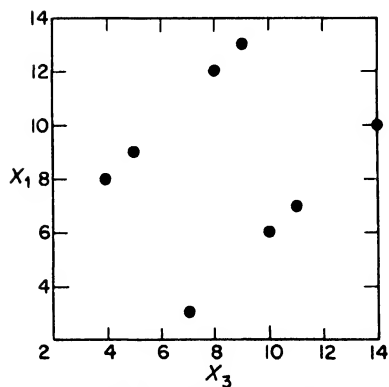
**CHART 21.2: TWO ARTIFICIAL TIME SERIES WITH IDENTICAL TRENDS, BUT CYCLES NEGATIVELY CORRELATED.**



(a) Original data



(b) Cycles



(c) Scatter diagram of original data

the series had been over a longer period of time, with the same tendencies in operation, the trend would have assumed even greater importance compared with the cycles, and the correlation coefficient would have had a still larger positive value.

## 21.2 FORECASTING A SERIES BY ITSELF

Most methods of forecasting a series by itself are naive in that they are lacking in theoretical basis.

**Synthesis of Projected Time Series Components.** If we can project the  $T$ ,  $C$ , and  $S$  components in the model

$$Y = TCSI$$

we can synthesize  $Y/I$  by multiplication of the  $T$ ,  $C$ , and  $S$  components. We cannot, of course, project irregular movements, because these movements are not systematic.

Projection of the trend is quite simple but not without hazard. The trend may be extended by the simple expedient of evaluating the trend equation for the desired  $X$  value or values in the future.

Trend projection for one year even on the basis of only 10 years is not unreasonable, but a projection becomes increasingly hazardous as the trend is projected further and further into the future. These difficulties should be kept in mind.

1. The economic causes affecting the trend may change after the type of trend equation has been determined.

2. The type of equation selected may not have been correct. There are many types of equations, and one of the bases of choosing among them is the way the trend behaves upon extension. Thus the trend is not used to forecast, but the forecast is used to select a trend.

3. The trend constants and, therefore, the trend values are subject to sampling error. The error in the trend values is a function of  $|x|$ , getting progressively larger as the trend is extended further from the chronological center of the data.

If the seasonal movement has been determined to be a stable one, a projection of the seasonal index into the future is quite simple. For example, if the stable seasonal index number for all Januaries in the past was determined to be 95.9, the seasonal index number for a forecasted January would probably be taken to be 95.9. If the seasonal index is moving, the projection of a future seasonal index number for a given month is sometimes accomplished by extension of a polynomial which has been used to smooth the  $SI$  values in obtaining the  $S$  values. Another common method is based on the assumption that  $S_{t+1} - S_t = 0.5(S_t - S_{t-1})$ , where  $S_t$  is the seasonal index number

for the current time period,  $S_{t+1}$  is the seasonal index number for one time period in the future, and  $S_{t-1}$  is the seasonal index number for one time period in the past. Projection of the cyclical movements is the crucial part of the method of projection. Not only is it the most difficult component to project, but it is the most important one to do accurately. The other methods considered in this chapter have to do primarily with forecasting cyclical movements. It is to be noted that the projection of the cycle is not purely mechanical but involves a considerable amount of judgment. Such projections should be made for forecasting purposes only by one who is thoroughly familiar with the business under consideration, and then only after careful analysis.

**Autoregression and Autocorrelation.** One of the characteristics of most economic time series is that successive observations are interdependent; there is correlation between the values at time  $t$  and time  $t - 1$ , perhaps also between values at time  $t$  and time  $t - 2$ , and so on. If  $Y$  represents values of the series to be forecasted, the regression equation might be

$$\hat{Y}_t = a + bY_{t-1} + cY_{t-2} \quad (21-1)$$

where  $\hat{Y}_t$  is the estimated value of  $Y$  at time  $t$ ,  $Y_{t-1}$  is the value of  $Y$  at time  $t - 1$ , and  $Y_{t-2}$  is the value of  $Y$  at time  $t - 2$ . Models like Eq. (21-1) are said to be autoregressive. The methods of Chapter 17 may be used to obtain the estimating equation. However, the method of least squares will give biased estimates of the coefficients of Eq. (21-1) under most conditions. It can be shown, however, that the bias decreases with increases in the sample size. Ordinary least squares should be used with caution when one is estimating the coefficients of models such as that illustrated by Eq. (21-1).<sup>(3)</sup>

Along these same lines *autocorrelation* coefficients may be calculated by using the model of Eq. (21-1) or a similar model.<sup>(4)</sup> If the model

$$\hat{Y}_t = a + bY_{t-1} \quad (21-2)$$

were estimated and a correlation coefficient calculated by the usual method, the correlation coefficient may be called a simple autocorrelation coefficient of lag 1, since  $Y_t$  is lagged behind itself one observation. Thus, if the  $Y_t$  series were

$t$	0	1	2	3	4	5
$Y_t$	1	3	6	8	7	5

<sup>(3)</sup> A discussion of ordinary least squares bias in models such as this can be found in J. Johnston, *Econometric Methods* (New York: McGraw-Hill Book Company, 1960), Chapter 8.

<sup>(4)</sup> Some writers prefer to use the term *serial correlation* coefficient when the time series is of finite length. However, modern usage tends to interchange the terms. Also, there are several other ways to define an autocorrelation coefficient. An enumeration of these other definitions is beyond the scope of this text.



the autocorrelation coefficient of lag 1 would be calculated by correlating the pairs

$t$	1	2	3	4	5
$Y_{t-1}$	1	3	6	8	7
$Y_t$	3	6	8	7	5

and an autocorrelation of lag 2 would be calculated by correlating the pairs

$t$	2	3	4	5
$Y_{t-2}$	1	3	6	8
$Y_t$	6	8	7	5

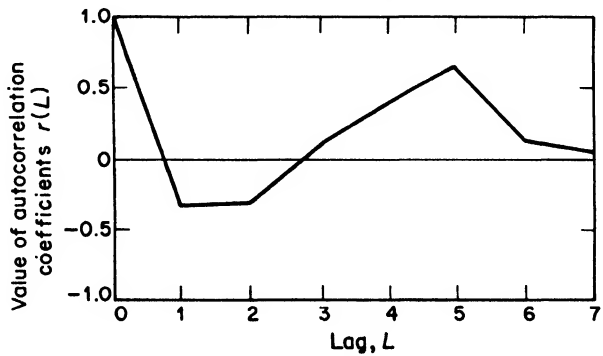
and so on for greater numbers of lags. Notice that if there are  $n$  observations on  $Y_t$ , a maximum of  $n - 2$  simple autocorrelation coefficients, with meaning, can be calculated, since by that time all but three observations on  $Y_t$  will be exhausted. If the autocorrelation coefficients are plotted against the lag  $L$  used in their calculation, a graph known as a *correlogram* results. Chart 21.3 shows a hypothetical correlogram. Apparently the highest correlation between the series and itself is when the lag is five time units, i.e., when the equation  $\hat{Y}_t = a + bY_{t-5}$  is used. Of course, the zero lag autocorrelation coefficient is unity.

Several warnings are in order when one is using autoregression as a forecasting device.

1. The standard deviation of the errors made in actual forecasting is likely to be larger than the computed standard error of estimate, since the estimating equation applies to an era that is past. It is unlikely that the same set of causes will persist in future years. In particular,

a. Consumer buying is affected by changes in the physical environment, by advertising appeals, and by the appearance of new products.

CHART 21.3: HYPOTHETICAL CORRELOGRAM.



b. Changes in methods of production affect the reaction of business to any stimulus.

c. Changes in the social environment, such as business combinations, growth of pressure groups, and new legislation are constantly affecting business behavior.

2. One should not mistake association for causation. Unless the regression relationship has a causal basis, however, it is not likely to persist.

3. It is difficult to determine the degree of significance to attach to our results. In the first place  $F$  and  $t$  tests are based upon the assumption that the data being analyzed are random samples of independent items drawn from a larger population. In economic time series, however, neither of these conditions is fulfilled. Each value in an economic time series is affected by preceding items and, in turn, affects the value of succeeding items. The number of degrees of freedom, therefore, is fewer than the number of observations. Also, as indicated above, the economic order is progressively changing. Each period studied is a unique situation, not a random occurrence which might happen at any time, though it is possible to regard any given set of data as being a sample result of a given set of causes operating in a chance manner.

4. Because of the fact that the number of observations upon which an autocorrelation coefficient is based decreases, under our definition of an autocorrelation coefficient, as the lag increases, the coefficients become less reliable as the lag approaches  $n - 3$ . Therefore, some statisticians offer the rule of thumb that autocorrelation coefficients should not be considered which are based upon a lag greater than 20 to 30 percent of the value of  $n$ . With 100 observations, not more than 30 autocorrelation coefficients should be calculated under this rule of thumb.

**Diffusion Indexes.** Let us consider a composite time series that is the sum or average of a large number of components. Usually all of the components will not reach cyclical peaks at the same time, nor will they reach cyclical troughs at the same time. On the other hand, the cyclical turning points will not be distributed at random over time, but most of the cyclical peaks will occur within a moderately short period of time, say a year, and most of the cyclical troughs also will occur within a moderately short period of time. This sort of behavior is perhaps inherent in a business economy where industries are interdependent and the effects of any stimulus are gradually diffused over the entire economy.

Under the conditions stated, when approximately one-half of the components have turned upward from their respective cyclical troughs, the composite series will also turn upward. Similarly, when approximately one-half of the components have turned downward, the composite series will also turn downward. Therefore, a *diffusion index* obtained by computing for

each month the percentage of the number of series that are expanding will lead the composite series. An index of this type is usually referred to as a *historical* diffusion index. The reason it is called a historical index is because the turning points of each series are determined historically. Only after watching a series for several months can we have much basis for distinguishing a cyclical turning point from a random or other irregular fluctuation. This "recognition lag" is the chief weakness of the historical diffusion index as a forecaster. By the time we know that a series has reached a cyclical turning point it is too late to use it as a forecaster.

Much experimentation has been made with *current* diffusion indexes. The simplest type is to count the number of series that are higher than the preceding month and then express this number as a percentage of all series. In general, current diffusion indexes are so irregular in appearance that they are difficult to use. Also, such an index typically ignores the relative magnitude of the change of different series and the relative importance of the different components. Such an index also typically ignores the reliability of the different components.

**Amplitude Adjusted Index.** One of the recent improvements in the reporting of current movements in important indicators of economic activity is the "amplitude adjusted" index developed by Julius Shiskin and Geoffrey Moore.<sup>(5)</sup> The index is composed of a weighted average of a group of component indicators. Each indicator is expressed in terms of standardized (amplitude-adjusted) percentage changes. Thus, for a given month, if a component has a standardized percentage increase of 3.0, the indicator is rising three times as fast as its average rate of change in the past. Each standardized component included in the resulting index has an equal opportunity to influence the index apart from the weight given it. This weight is determined by the forecasting value of the component according to National Bureau of Economic Research criteria.

**Percentage Change in a Time Series.** An economic time series does not ordinarily increase by a constant amount or percentage until it reaches its peak and then decline by a constant amount or percentage until it reaches its trough. Rather, there is some tendency to slacken its rate of growth or decline before it reaches a turning point. Therefore, the percentage change in a time series will tend to change direction before the series itself. It might be thought that such a series of percentage changes would be useful for forecasting purposes. The difficulty is that the irregularities of a time series are accentuated by taking percentage changes. Consequently, the

---

<sup>(5)</sup> See Julius Shiskin and Geoffrey H. Moore, *Composite Indexes of Leading, Coinciding and Lagging Indicators, 1948-67*, Supplement to National Bureau Report 1, The National Bureau of Economic Research, New York, January, 1968.

series of percentage changes is very irregular in appearance and hard to interpret.

### 21.3 FORECASTING A SERIES BY OTHER SERIES

Since the cyclical turning points in all series do not occur at the same time, it is sometimes possible to predict the turning points or even the values of one series if one or more forecasting series can be found that precede with some degree of regularity the one we wish to predict. There are various ways of estimating leads and lags of different series. The same results are not necessarily obtained by the different methods.

A very simple model, which uses a related series  $X$  to forecast the series  $Y$ , is

$$\hat{Y}_t = a + bX_t$$

Also, the independent variable may be lagged

$$\hat{Y}_t = a + bX_{t-1}$$

and *cross-correlation* coefficients calculated, i.e., the simple correlation coefficients for each lag, as an aid in the determination of the optimal lag for  $X_t$ . Of course, other kinds of lag relationships are possible. For example,

$$Y_t = a + bX_t + cX_{t-1} + dX_{t-2} + \dots$$

and so on.<sup>(6)</sup> Again, the objections and difficulties stated in the last section to the use of autocorrelation coefficients apply to the use of cross-correlation coefficients. In addition, the following difficulties should be noted:

1. The degree of correlation obtained between two series depends on the nature of the trend around which the cyclical movements are measured. The timing of the cyclical movements may be similar, after adjustment for lag, but the relative amplitude of the deviations of the two series may vary over different segments of time. This difficulty may sometimes be overcome by using a more flexible trend. Another device is to correlate either differences between one month and the preceding or percentages of the preceding month. When this is done, the data are not previously adjusted for trend, since these methods partially eliminate the trend.

2. It is sometimes hard to tell which series is the forecaster. Although series  $A$  may precede series  $B$  at recessions, series  $B$  may typically precede series  $A$  at revivals. In other cases series  $A$ , when lagged, may precede series

<sup>(6)</sup> Models of this type are often referred to as "distributed lag" models. Since the successive values of  $X$  are usually highly correlated, estimation of this model by the method of ordinary least squares is not to be recommended. For an interesting approach to estimation, see: L. M. Koyck, *Distributed Lags and Investment Analysis* (Amsterdam: North-Holland Publishing Company, 1954).

*B* with positive correlation, but series *B*, when lagged, may precede series *A* with a negative correlation.<sup>(7)</sup> Thus expanding business may bring about higher interest rates, but high interest rates may bring about a business decline.

3. Often the series we wish to forecast is one that moves early in the business cycle. Stock prices are an example of such a series.

4. Another factor that impairs the usefulness of this method is the scarcity of time series on a basis shorter than a month. It is quite possible that weekly, daily, or hourly data might bring to light relationships that are known and utilized only by a few "insiders." It does not seem logical that the cause-and-effect relationships that supposedly surround us on every side must take a month or more for their development. There must be many that work out in a few days, a few hours, or nearly instantaneously. As data are made available upon a weekly, daily, or more frequent basis, it is conceivable that very reliable lags and leads may be obtained. These may assist in accurate forecasting and improved control of business processes by the businessman.

## 21.4 SPECIFIC HISTORICAL ANALOGY

Since all cycles are not uniform in amplitude or duration, some forecasters make use of history, not by projecting any fancied economic rhythm into the future or relying on any repetitive sequence, but by selecting some specific previous situation which has many of the earmarks of the present and concluding that what happened in that previous situation will happen in the present one. As of the summer of 1945, a favorite analogy was that of the period following World War I. It was pointed out that there was a short demobilization depression in 1919, a restocking boom in 1920 followed by a sharp deflation trough in 1921, and that business was carried forward by a revival of the durable goods industry in ensuing years. In the early part of 1957 the inflationary tendencies brought to some people's minds the fact that there were inflationary tendencies (though not entirely of the same type) also in 1929, and that they culminated in a stock market crash and business collapse.

Although it is undoubtedly true that partial analogies can be discovered in past history, one should be careful to take into consideration the differences as well as the similarities between past and present situations. The differences in the amount and type of government intervention in economic affairs should be especially noted.

---

<sup>(7)</sup> It is even conceivable that a series representing the effect will precede a series representing the cause. Thus stock prices are a function of earnings and dividends, but in evaluating a stock one usually strongly considers his expectation of future earnings and dividends.

## 21.5 SURVEYS OF PLANS AND OPINIONS

Some idea of prospects for future months can often be obtained by the analysis of questionnaires designed to reveal the plans and/or opinions of economists, business executives, or consumers. Among organizations using this technique are *Fortune* Magazine, United States Department of Commerce, Securities and Exchange Commission, and University of Michigan Survey Research Center.

Although considered helpful for forecasting purposes, surveys of inclination to buy have their usefulness somewhat diminished by two factors. First, a fairly large proportion of the respondents apparently do not take their answers seriously. Second, consumer plans are usually not very firm and may be changed rather quickly.

It is not necessarily the case that best results are obtained by a representative sample. When opinions are sampled, the sample should contain only those who are well informed and whose opinions are sound. In general, executives from large firms have means of obtaining better advice than do those of small firms. When the plans of businessmen and consumers are sampled, a sample of people who will carry out their plans and who have already made commitments is better than one of people who do not know what they will do or who are likely to change their minds.

Some polls of opinions contain many answers that are uninformed, careless guesses. If included, such opinions tend to average out. There is evidence that many of the opinions are based on simple projections of recent patterns. The cyclical peaks and troughs are, therefore, not anticipated. Another tendency is to underestimate the extent of a cyclical rise or a cyclical fall.

## 21.6 CROSSCUT ECONOMIC ANALYSIS

This method is based upon the theory that no two cycles are alike but that like causes always produce like results. All the factors bearing upon a given situation are assembled, and, relying upon his knowledge of economic processes, the forecaster concludes whether the situation is favorable or unfavorable. Although the method is essentially nonstatistical, it is possible to develop a statistical technique by assigning weights to each factor and then counting the score to see whether the net result is favorable or unfavorable.

At the middle of 1957, one of the authors counted six favorable factors mentioned in business periodicals and 11 unfavorable factors. Business economists were somewhat baffled by the mixed indications. However, some were impressed with the theory that the many built-in or managed stabilizing devices would avert a serious depression. The words "rolling adjustment"

and "sideways movement" were in vogue. The majority opinion seemed to be that there would be little change in the fortunes of business until the last quarter, when most time series would show improving business! This view may perhaps be attributed to professional optimism and partly to a psychological lag.

## 21.7 MULTIPLE EQUATION MODELS

In many situations a single equation is not adequate for specifying the behavior of a variable of interest. Further, there are instances in which it is desirable to forecast the behavior of several variables simultaneously. Simultaneous equation models are, therefore, often employed as a forecasting device.

**Causal Models.** Suppose that we believe that inventory holding in this time period  $I_t$  depends in a linear fashion on the price of the goods held in the last time period  $P_{t-1}$ . Then

$$I_t = a_1 + b_1 P_{t-1}$$

Also, suppose that we specify that the price of the goods held in this time period depends upon the inventory of those goods currently being held.

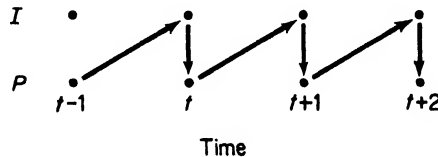
$$P_t = a_2 + b_2 I_t$$

Of course, we could include other variables in the model, but, for simplicity, we will not include these variables in the discussion. The complete model

$$I_t = a_1 + b_1 P_{t-1}$$

$$P_t = a_2 + b_2 I_t$$

is said to be a pure causal chain. The diagram below, sometimes called a *Tinbergen arrow diagram*, shows that the value of  $P$  in the last time period determines the value of  $I$  in the current time period, which in turn determines the value of  $P$  in the current time period.



Pure causal chain models can be estimated by applying the method of least squares to each equation of the model in the manner familiar to the student. Models of this type are sometimes called *recursive*.

**Interdependent Models.** Suppose that the quantity of a commodity supplied in this time period  $Q_t$  is a function of the price of the commodity in this time period  $P_t$  and the cost of production in this time  $C_t$ . Then

$$\text{Supply: } Q_t = a_1 + b_1 P_t + c_1 C_t$$

Furthermore, suppose that the quantity demanded  $Q_t$  (which must be the same as the quantity supplied in market equilibrium) is given by

$$\text{Demand: } Q_t = a_2 + b_2 P_t + c_2 Y_t$$

where  $Y_t$  is national income in the current time period. Together, the demand and supply equation form the *structural equations* of an interdependent model. In models of this type the direct application of least squares to each equation will generally lead to biased estimates of the parameters, since  $Q_t$  and  $P_t$  are *jointly determined* variables. The variables  $C_t$  and  $Y_t$  are said to be *predetermined* variables, since their values are given outside the model. In some cases we can write models of this type in a form such that only one jointly determined variable appears in any equation. For example, from the supply equation

$$P_t = \frac{Q_t - a_1 - c_1 C_t}{b_1}$$

and substituting  $P_t$  into the demand equation, we find that

$$Q_t = a_2 + \frac{b_2}{b_1} (Q_t - a_1 - c_1 C_t) + c_2 Y_t$$

$$\text{or} \quad Q_t \left( 1 - \frac{b_2}{b_1} \right) = a_2 - \frac{b_2}{b_1} a_1 - \frac{b_2}{b_1} c_1 C_t + c_2 Y_t$$

If we let

$$a = \frac{a_2 - \frac{b_2}{b_1} a_1}{1 - \frac{b_2}{b_1}} \quad b = \frac{-\frac{b_2}{b_1} c_1}{1 - \frac{b_2}{b_1}}$$

and

$$c = \frac{c_2}{1 - \frac{b_2}{b_1}}$$

We can write the demand equation as

$$Q_t = a + b C_t + c Y_t \quad (21-3)$$

which contains only one jointly dependent variable. Similarly, substituting  $Q_t$  as defined by Eq. (21-3) into the supply equation, we have

$$P_t = a^* + b^* C_t + c^* Y_t \quad (21-4)$$



where

$$a^* = \frac{a - a_1}{b_1}$$

$$b^* = \frac{b - c_1}{b_1}$$

$$c^* = \frac{c}{b_1}$$

When the structural equations of a model are written so that each equation contains one and only one jointly determined variable, the model is said to be written in *reduced form*. Equations (21-3) and (21-4) are, therefore, the reduced form of our demand and supply model. In general, the parameters of Eqs. (21-3) and (21-4) may be estimated by the method of ordinary least squares. Since there is only one jointly determined variable per equation, the estimates of the parameters will be unbiased under the assumptions set out in the appendix to Chapter 16. The estimation method is known as the method of *indirect least squares*. The final step consists of solving back into the structural equations to find  $a_1$ ,  $b_1$ ,  $c_1$ ,  $a_2$ ,  $b_2$ , and  $c_2$ . For example, from Eq. (21-3) we can estimate  $c$  and from Eq. (21-4) we can estimate  $c^*$ . Then, since  $c^* = c/b_1$ , we can solve for  $b_1$ , and so on. If the estimates of the parameters of the structural equations can be uniquely deduced from the estimates of the reduced form parameters, the model is said to be *exactly identified*.

## 21.8 JUDGING THE ACCURACY OF A FORECAST

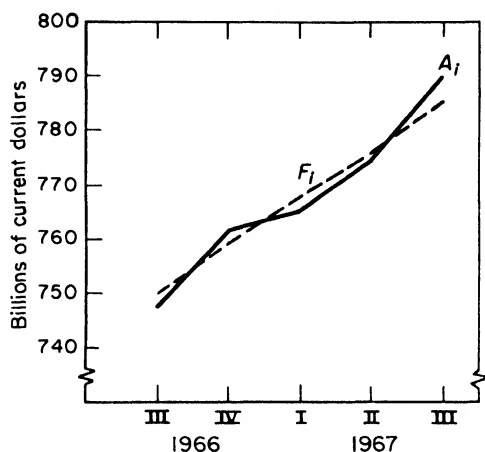
Once a forecast of a time series has been made, a natural question which arises is, "How good is the forecast?" Of course, the question might be stated, "Over the forecast period, how close is the correspondence between the forecasted values of the time series and the actual values of the time series?" There are many ways to approach these questions, but we will review only one, Theil's coefficient.<sup>(8)</sup> Following Theil, let us define the coefficient

$$U = \frac{\sqrt{\frac{1}{n} \sum (F_i - A_i)^2}}{\sqrt{\frac{1}{n} \sum F_i^2} + \sqrt{\frac{1}{n} \sum A_i^2}} \quad (21-5)$$

where the series  $F_i$  represents the forecasted values of a time series and  $A_i$ , the actual values of the time series. If the predicted and actual values are

<sup>(8)</sup> H. Theil, *Economic Forecasts and Policy* (Amsterdam: North-Holland Publishing Company, 1958). An alternative coefficient is found in Herman O. A. Wold, ed., *Econometric Model Building* (Amsterdam: North-Holland Publishing Company, 1964), Chapter 5.

**CHART 21.4: FORECASTED AND ACTUAL VALUES OF GNP, UNITED STATES, THIRD QUARTER 1966 TO THIRD QUARTER 1967.**



*Source: Actual values from Federal Reserve Bulletin (November 1967). Forecasted values are hypothetical.*

identical in the prediction interval, then the numerator of  $U$  will be zero. Thus, a "perfect" forecast is obtained when  $U = 0$ . On the other hand, the largest value  $U$  can assume is one (except in the case where both  $A_t$  and  $F_t$  are zero) and, in this case, it may be argued that the forecast is as "bad" as it can possibly be. Specifically,  $U$  will be zero when all  $F_t = A_t$ . The coefficient will be one when  $F_t = -A_t$  or when all  $F_t$  values or all  $A_t$  values are zero (but not both).  $U^2$  is analogous to a coefficient of nondetermination, since it is a measure of nonagreement between the actual and forecasted values, and because  $U$  lies in a fixed and convenient interval.

$$0 \leq U \leq 1$$

However, as we shall see, the  $U$  coefficient has several advantages over a simple coefficient of nondetermination as an index of forecast inaccuracy.

To illustrate, suppose that gross national product in the United States had been forecasted by the dotted line in Chart 21.4. The actual values are shown by the solid line. Table 21.1 shows the actual and forecasted values of the time series, and from it we obtain

$$U = \frac{\sqrt{\frac{34.36}{5}}}{\sqrt{\frac{2,953,752}{5}} + \sqrt{\frac{2,949,930}{5}}}$$

$$U = \frac{2.62}{1536.70} = 0.00171$$

**TABLE 21.1: ACTUAL AND FORECASTED VALUES OF GNP, UNITED STATES, THIRD QUARTER 1966 TO THIRD QUARTER 1967, BILLIONS OF CURRENT DOLLARS AND CALCULATION OF THEIL'S  $U$**

Date	Actual $A_i$	Forecasted $F_i$	$F_i - A_i$	$(F_i - A_i)^2$	$A_i^2$	$F_i^2$
1966: III	748.8	750.0	1.2	1.44	560,701.44	562,500.00
IV	762.1	759.0	-3.1	9.61	580,796.41	576,081.00
1967: I	766.3	768.0	1.7	2.89	587,215.69	589,824.00
II	775.1	777.0	1.9	3.61	600,780.01	603,729.00
III	790.1	786.0	-4.1	16.81	624,258.01	617,796.00
Total	3842.4	3840.0	...	34.36	2,953,751.56	2,949,930.00

The value of Theil's coefficient is close to zero, and this seems to indicate a rather accurate forecast.<sup>(9)</sup> Certainly the  $U$  coefficient could be used to compare this forecast with an alternative forecast.

One very interesting aspect of Theil's  $U$  is that it can be decomposed into three percentage components: a component attributable to lack of equality of the arithmetic means of  $A_i$  and  $F_i$ , a part attributable to the lack of equality in the standard deviations of the samples of  $A_i$  and  $F_i$ , and a part attributable to the lack of correlation between  $A_i$  and  $F_i$ . Thus<sup>(10)</sup>

$$U^2 = U_1^2 + U_2^2 + U_3^2$$

$$= \left( \frac{F - \bar{A}}{D} \right)^2 + \left( \frac{SD_F - SD_A}{D} \right)^2 + \frac{2(1 - r)SD_F SD_A}{D^2} \quad (21-6)$$

where  $D$  is the denominator of Eq. (21-5),  $F$  and  $\bar{A}$  are the arithmetic means of  $F_i$  and  $A_i$ , and  $SD_F$  and  $SD_A$  are the standard deviations of the sample of  $F_i$  and  $A_i$ . Multiplying both sides of Eq. (21-6) by  $100/U^2$ , we see that we may decompose  $U^2$  into percentage components. For example, from Table 21.1

$$U_1^2 = \left( \frac{F - \bar{A}}{D} \right)^2 = \left( \frac{768 - 768.48}{1536.70} \right)^2 = (-0.0003124)^2$$

so that the percentage of  $U$  attributable to differences between the means of  $F_i$  and  $A_i$  is

$$(-0.0003124)^2 \left[ \frac{100}{(0.00171)^2} \right] = 3$$

<sup>(9)</sup> Actually, one cannot judge the accuracy of an individual forecast by looking at  $U$  alone. Under rather strict assumptions a standard error can be developed for  $U$ . See Theil, *op. cit.*, pp. 43ff.

<sup>(10)</sup> The proof that  $U^2 = U_1^2 + U_2^2 + U_3^2$  is straightforward if one recalls

that

$$(SD_{F-A})^2 = (SD_F)^2 + (SD_A)^2 - 2rSD_F SD_A$$

$$= \frac{\sum (F - A)^2}{n} - \left( \frac{\sum F - \sum A}{n} \right)^2$$

Similarly, the student may verify that  $SD_F = 12.73$  and  $SD_A = 13.74$  so that

$$U_{\frac{1}{2}}^2 = \left( \frac{12.73 - 13.74}{1536.70} \right)^2 = (-0.0006573)^2$$

and the percentage of  $U$  attributable to differences between the standard deviations of  $F_t$  and  $A_t$  is

$$(-0.0006573)^2 \left[ \frac{100}{(0.00171)^2} \right] = 15$$

which means that the remaining 82 percent is attributable to lack of correlation between the two series. Presumably the greatest room for improvement in the forecasting technique lies in the area of attempting to increase the correlation between the series.

## PROBLEMS

1. Suppose that the series  $Y_t$  repeats itself every four time units; i.e., has a cycle with period 4. What will the correlogram of this series look like if we have sufficient observations on  $Y_t$ ?

2. The following data represent percentage deviations from trend for current construction expenditures and current sales of the Union Carbide and Carbon Company for several years:

*a. Estimate expenditures, using sales. (Lag the relationship if you think that it will help.) Forecast expenditures for the next year.*

Current expenditures	12	18	-30	-35	1	29	31	-12	-20	6
Current sales	3	7	-14	-1	10	3	1	-16	4	4

3. In the supply and demand model of Sec. 21.7 suppose that ordinary least squares applied to data for Eqs. (21-3) and (21-4) gave the following estimates:

$$\begin{aligned} a &= 1 & a^* &= 4 \\ b &= 2 & b^* &= 5 \\ c &= 3 & c^* &= 6 \end{aligned}$$

What are the derived estimates of the parameters of the structural equations?

4. Prove Eq. (21-6).



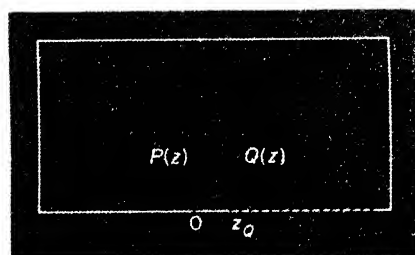
## Appendixes



# APPENDIX 1: Values of $Q(z)$ for Selected Values of $z_Q$

$z$	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.50000	0.49601	0.49202	0.48803	0.48405	0.48006	0.47608	0.47210	0.46812	0.46414
0.1	0.46017	0.45620	0.45224	0.44828	0.44433	0.44038	0.43644	0.43251	0.42858	0.42465
0.2	0.42074	0.41683	0.41294	0.40905	0.40517	0.40129	0.39743	0.39358	0.38974	0.38591
0.3	0.38209	0.37828	0.37448	0.37070	0.36693	0.36317	0.35942	0.35569	0.35197	0.34827
0.4	0.34458	0.34090	0.33724	0.33360	0.32997	0.32636	0.32276	0.31918	0.31561	0.31207
0.5	0.30854	0.30503	0.30153	0.29806	0.29460	0.29116	0.28774	0.28434	0.28096	0.27760
0.6	0.27425	0.27093	0.26763	0.26435	0.26109	0.25785	0.25463	0.25143	0.24825	0.24510
0.7	0.24196	0.23885	0.23576	0.23270	0.22965	0.22663	0.22363	0.22065	0.21770	0.21476
0.8	0.21186	0.20897	0.20611	0.20327	0.20045	0.19766	0.19489	0.19215	0.18943	0.18673
0.9	0.18406	0.18141	0.17879	0.17619	0.17361	0.17106	0.16853	0.16602	0.16354	0.16109
1.0	0.15866	0.15625	0.15386	0.15151	0.14917	0.14686	0.14457	0.14231	0.14007	0.13786
1.1	0.13567	0.13350	0.13136	0.12924	0.12714	0.12507	0.12302	0.12100	0.11900	0.11702
1.2	0.11507	0.11314	0.11123	0.10935	0.10749	0.10565	0.10383	0.10204	0.10027	0.09853
1.3	0.09680	0.09510	0.09342	0.09176	0.09012	0.08851	0.08694	0.08534	0.08379	0.08226
1.4	0.08076	0.07927	0.07780	0.07636	0.07493	0.07353	0.07215	0.07078	0.06944	0.06811
1.5	0.06681	0.06552	0.06426	0.06301	0.06178	0.06057	0.05938	0.05821	0.05705	0.05592
1.6	0.05480	0.05370	0.05262	0.05155	0.05050	0.04947	0.04846	0.04746	0.04648	0.04551
1.7	0.04457	0.04363	0.04272	0.04182	0.04093	0.04006	0.03920	0.03836	0.03754	0.03673
1.8	0.03593	0.03515	0.03438	0.03362	0.03288	0.03216	0.03144	0.03074	0.03005	0.02938
1.9	0.02872	0.02807	0.02743	0.02680	0.02619	0.02559	0.02500	0.02442	0.02385	0.02330
2.0	0.02275	0.02216	0.02169	0.02121	0.02068	0.02018	0.01970	0.01923	0.01876	0.01831
2.1	0.01786	0.01743	0.01700	0.01659	0.01618	0.01578	0.01539	0.01500	0.01463	0.01426
2.2	0.01390	0.01355	0.01321	0.01287	0.01255	0.01222	0.01191	0.01160	0.01130	0.01101
2.3	0.01072	0.01044	0.01017	0.00990	0.00964	0.00939	0.00914	0.00889	0.00866	0.00842
2.4	0.00820	0.00798	0.00776	0.00755	0.00734	0.00714	0.00695	0.00676	0.00657	0.00639
2.5	0.00621	0.00604	0.00587	0.00570	0.00554	0.00539	0.00523	0.00508	0.00494	0.00480
2.6	0.00466	0.00453	0.00440	0.00427	0.00415	0.00402	0.00391	0.00379	0.00368	0.00357
2.7	0.00347	0.00336	0.00326	0.00317	0.00307	0.00298	0.00289	0.00280	0.00272	0.00264
2.8	0.00256	0.00248	0.00240	0.00233	0.00226	0.00219	0.00212	0.00205	0.00199	0.00193
2.9	0.00187	0.00181	0.00175	0.00169	0.00164	0.00159	0.00154	0.00149	0.00144	0.00139
3.0	0.00135	0.00131	0.00126	0.00122	0.00118	0.00114	0.00111	0.00107	0.00104	0.00100
3.1	0.00097	0.00094	0.00090	0.00087	0.00084	0.00082	0.00079	0.00076	0.00074	0.00071
3.2	0.00069	0.00066	0.00064	0.00062	0.00060	0.00058	0.00056	0.00054	0.00052	0.00050
3.3	0.00048	0.00047	0.00045	0.00043	0.00042	0.00040	0.00039	0.00038	0.00036	0.00035
3.4	0.00034	0.00032	0.00031	0.00030	0.00029	0.00028	0.00027	0.00026	0.00025	0.00024
3.5	0.00023	0.00022	0.00022	0.00021	0.00020	0.00019	0.00019	0.00018	0.00017	0.00017
3.6	0.00016	0.00015	0.00015	0.00014	0.00014	0.00013	0.00013	0.00012	0.00012	0.00011
3.7	0.00011	0.00010	0.00010	0.00010	0.00009	0.00009	0.00008	0.00008	0.00008	0.00008
3.8	0.00007	0.00007	0.00007	0.00006	0.00006	0.00006	0.00006	0.00005	0.00005	0.00005
3.9	0.00005	0.00005	0.00004	0.00004	0.00004	0.00004	0.00004	0.00004	0.00003	0.00003
4.0	0.00003	..	..	..	..	..	..	..	..	..
4.5	0.000003	..	..	..	..	..	..	..	..	..
5.0	0.0000003	..	..	..	..	..	..	..	..	..

$$Q(z) = \int_z^{\infty} f(u) du; \quad P(z) = 1 - Q(z)$$

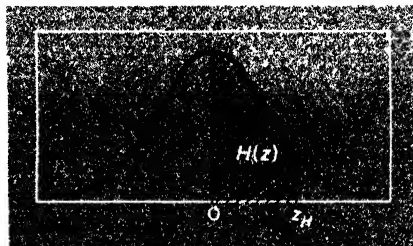




## APPENDIX 2: Values of $H(z)$ for Selected Values of $z_H$

$z$	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.00000	0.00399	0.00798	0.01197	0.01595	0.01994	0.02392	0.02790	0.03188	0.03586
0.1	0.03983	0.04380	0.04776	0.05172	0.05567	0.05962	0.06356	0.06749	0.07142	0.07535
0.2	0.07926	0.08317	0.08706	0.09095	0.09483	0.09871	0.10257	0.10642	0.11026	0.11409
0.3	0.11791	0.12172	0.12552	0.12930	0.13307	0.13683	0.14058	0.14431	0.14803	0.15173
0.4	0.15542	0.15910	0.16276	0.16640	0.17003	0.17364	0.17724	0.18082	0.18439	0.18793
0.5	0.19146	0.19497	0.19847	0.20194	0.20540	0.20884	0.21226	0.21566	0.21904	0.22240
0.6	0.22575	0.22907	0.23237	0.23565	0.23891	0.24215	0.24537	0.24857	0.25175	0.25490
0.7	0.25804	0.26115	0.26424	0.26730	0.27035	0.27337	0.27637	0.27935	0.28230	0.28524
0.8	0.28814	0.29103	0.29389	0.29673	0.29955	0.30234	0.30511	0.30785	0.31057	0.31327
0.9	0.31594	0.31859	0.32121	0.32381	0.32639	0.32894	0.33147	0.33398	0.33646	0.33891
1.0	0.34134	0.34375	0.34614	0.34849	0.35083	0.35314	0.35543	0.35769	0.35993	0.36214
1.1	0.36433	0.36650	0.36864	0.37076	0.37286	0.37493	0.37698	0.37900	0.38100	0.38298
1.2	0.38493	0.38686	0.38877	0.39065	0.39251	0.39435	0.39617	0.39796	0.39973	0.40147
1.3	0.40320	0.40490	0.40658	0.40824	0.40988	0.41148	0.41309	0.41466	0.41621	0.41774
1.4	0.41924	0.42073	0.42220	0.42364	0.42507	0.42647	0.42785	0.42922	0.43056	0.43189
1.5	0.43319	0.43448	0.43574	0.43699	0.43822	0.43943	0.44062	0.44179	0.44295	0.44408
1.6	0.44520	0.44630	0.44738	0.44845	0.44950	0.45053	0.45154	0.45254	0.45352	0.45449
1.7	0.45543	0.45637	0.45728	0.45818	0.45907	0.45994	0.46080	0.46164	0.46246	0.46327
1.8	0.46407	0.46485	0.46562	0.46638	0.46712	0.46784	0.46856	0.46926	0.46995	0.47062
1.9	0.47128	0.47193	0.47257	0.47320	0.47381	0.47441	0.47500	0.47558	0.47615	0.47670
2.0	0.47725	0.47784	0.47831	0.47882	0.47932	0.47982	0.48030	0.48077	0.48124	0.48169
2.1	0.48214	0.48257	0.48300	0.48341	0.48382	0.48422	0.48461	0.48500	0.48537	0.48574
2.2	0.48610	0.48645	0.48679	0.48713	0.48745	0.48778	0.48809	0.48840	0.48870	0.48899
2.3	0.48928	0.48956	0.48983	0.49010	0.49036	0.49061	0.49086	0.49111	0.49134	0.49158
2.4	0.49180	0.49202	0.49224	0.49245	0.49266	0.49286	0.49305	0.49324	0.49343	0.49361
2.5	0.49379	0.49396	0.49413	0.49430	0.49446	0.49461	0.49477	0.49492	0.49506	0.49520
2.6	0.49534	0.49547	0.49560	0.49573	0.49585	0.49598	0.49609	0.49621	0.49632	0.49643
2.7	0.49653	0.49664	0.49674	0.49683	0.49693	0.49702	0.49711	0.49720	0.49728	0.49736
2.8	0.49744	0.49752	0.49760	0.49767	0.49774	0.49781	0.49788	0.49795	0.49801	0.49807
2.9	0.49813	0.49819	0.49825	0.49831	0.49836	0.49841	0.49846	0.49851	0.49856	0.49861
3.0	0.49865	0.49869	0.49874	0.49878	0.49882	0.49886	0.49889	0.49893	0.49896	0.49900
3.1	0.49903	0.49906	0.49910	0.49913	0.49916	0.49918	0.49921	0.49924	0.49926	0.49929
3.2	0.49931	0.49934	0.49936	0.49938	0.49940	0.49942	0.49944	0.49946	0.49948	0.49950
3.3	0.49952	0.49953	0.49955	0.49957	0.49958	0.49959	0.49960	0.49961	0.49962	0.49964
3.4	0.49966	0.49968	0.49969	0.49970	0.49971	0.49972	0.49973	0.49974	0.49975	0.49976
3.5	0.49977	0.49978	0.49978	0.49979	0.49980	0.49981	0.49981	0.49982	0.49983	0.49983
3.6	0.49984	0.49985	0.49985	0.49986	0.49986	0.49987	0.49987	0.49988	0.49988	0.49989
3.7	0.49989	0.49990	0.49990	0.49990	0.49991	0.49991	0.49992	0.49992	0.49992	0.49992
3.8	0.49993	0.49993	0.49993	0.49994	0.49994	0.49994	0.49994	0.49995	0.49995	0.49995
3.9	0.49995	0.49996	0.49996	0.49996	0.49996	0.49996	0.49996	0.49996	0.49997	0.49997
4.0	0.49997	...	...	...	...	...	...	...	...	...
4.5	0.499997	...	...	...	...	...	...	...	...	...
5.0	0.4999997	...	...	...	...	...	...	...	...	...

$$H(z) = \int_0^{z_H} f(u) du$$



### APPENDIX 3: Values of $z_Q$ for Selected Values of $Q(z)$

$Q(z)$	$z_Q$	$Q(z)$	$z_Q$	$Q(z)$	$z_Q$
0.0005	3.29053	0.005	2.57583	0.11	1.22653
0.0010	3.09023	0.010	2.32635	0.12	1.17499
0.0015	2.96774	0.015	2.17009	0.13	1.12639
0.0020	2.87816	0.020	2.05375	0.14	1.08032
0.0025	2.80703	0.025	1.95996	0.15	1.03643
0.0030	2.74778	0.030	1.88079	0.16	0.99446
0.0035	2.69684	0.035	1.81191	0.17	0.95417
0.0040	2.65207	0.040	1.75069	0.18	0.91537
0.0045	2.61205	0.045	1.69540	0.19	0.87790
0.0050	2.57583	0.050	1.64485	0.20	0.84162
0.006	2.51214	0.06	1.55477	0.25	0.67449
0.007	2.45726	0.07	1.47579	0.30	0.52440
0.008	2.40892	0.08	1.40507	0.35	0.38532
0.009	2.36562	0.09	1.34076	0.40	0.25335
0.010	2.32635	0.10	1.28155	0.45	0.12566

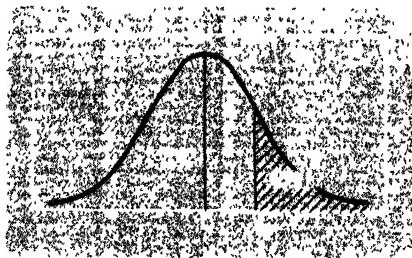
$$z = \frac{X - \mu}{\sigma} \quad \text{or} \quad \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}$$

#### Examples:

A probability of 0.025 is associated with a deviation from  $\mu$  as large as  $+1.9599\sigma$  or larger (area in upper tail is 0.025).

A probability of 0.025 is associated with a deviation from  $\mu$  as small as  $-1.9599\sigma$  or smaller (area in lower tail is 0.025).

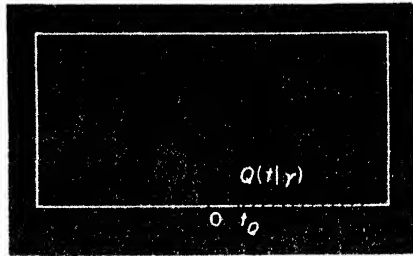
A probability of 0.05 is associated with a deviation from  $\mu$  as large numerically as 1.9599 or larger (area in both tails is 0.05).



APPENDIX 4: Values of  $t_Q$  for Selected Values of  $Q(t|\nu)$

$\nu$	0.45	0.40	0.35	0.30	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.158	0.325	0.510	0.727	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	636.692
2	0.142	0.289	0.445	0.617	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.598
3	0.137	0.277	0.424	0.584	0.765	0.978	1.250	1.638	2.312	3.182	4.941	5.841	12.924
4	0.134	0.271	0.414	0.569	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	0.132	0.267	0.408	0.559	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	0.131	0.265	0.404	0.553	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.130	0.263	0.402	0.549	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	0.130	0.262	0.399	0.546	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	0.129	0.261	0.398	0.543	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	0.129	0.260	0.397	0.542	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	0.129	0.260	0.396	0.540	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	0.128	0.259	0.395	0.539	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	0.128	0.259	0.394	0.538	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	0.128	0.258	0.393	0.537	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	0.128	0.258	0.393	0.536	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	0.128	0.258	0.392	0.535	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	0.128	0.257	0.392	0.534	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	0.127	0.257	0.392	0.534	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.877	3.922
19	0.127	0.257	0.391	0.533	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	0.127	0.257	0.391	0.533	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	0.127	0.257	0.391	0.532	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	0.127	0.256	0.390	0.532	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	0.127	0.256	0.390	0.532	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.767
24	0.127	0.256	0.390	0.531	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	0.127	0.256	0.389	0.531	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	0.127	0.256	0.389	0.530	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.659
30	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
40	0.126	0.255	0.388	0.529	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551
60	0.126	0.254	0.387	0.527	0.679	0.848	1.046	1.296	1.671	2.000	2.390	2.660	3.460
120	0.126	0.254	0.386	0.526	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617	3.373
$\infty$	0.126	0.253	0.385	0.524	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.291

This table is reprinted from Table III of R. A. Fisher and F. Yates, Statistical Tables for Biological, Agricultural and Medical Research, 5th ed.



**APPENDIX 5: Values of  $F_Q$  for Selected Values of  $Q(F | \nu_1, \nu_2)$**

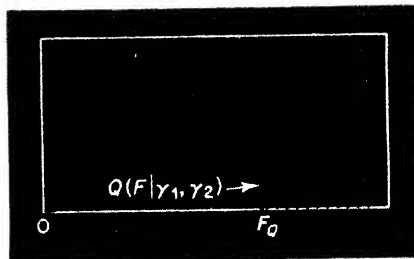
$$F = \frac{\chi_1^2/\nu_1}{\chi_2^2/\nu_2} = \frac{s_1^2}{s_2^2}$$

where  $s_1^2$  and  $s_2^2$  are two independent estimates of  $\sigma^2$ . The degrees of freedom for the numerator and denominator are  $\nu_1$  and  $\nu_2$ , respectively. The tables give upper probability points.

A lower probability point is the reciprocal of  $F$  with  $\nu_1$  and  $\nu_2$  interchanged. Thus if  $\nu_1 = 5$  and  $\nu_2 = 8$ , the 0.05 upper probability point is 3.69. If  $\nu_1 = 8$  and  $\nu_2 = 5$  the lower probability point is  $1/3.69 = 0.271$ .

To estimate values of  $F$  values of  $\nu_1$  or  $\nu_2$  not given in this table interpolate, using reciprocals of  $\nu$ .

Values of  $F$  at the 0.05 and 0.01 points were abridged, by permission, from E. S. Pearson and H. O. Hartley (editors), *Biometrika Tables for Statisticians*, Volume I, Cambridge University Press, Cambridge, 1954, Table 18. The *Biometrika* tables are more extensive with respect to  $\nu_1$  and  $\nu_2$ . Values of  $F$  at the 0.001 points were abridged from Table V of R. A. Fisher and F. Yates, *Statistical Tables for Biological, Agricultural, and Medical Research*, 5th ed., 1957, published by Oliver and Boyd, Ltd., Edinburgh, by permission.



VALUES OF  $F_Q$  FOR  $Q(F | v_1, v_2) = 0.05$ 

$v_2 \backslash v_1$	1	2	3	4	5	6	8	10	12	20	24	30	40	60	120	$\infty$
1	161.4	199.5	215.7	224.6	230.2	234.0	238.9	241.9	243.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.40	19.41	19.45	19.45	19.47	19.46	19.47	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.85	8.79	8.74	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.96	5.91	5.80	5.77	5.72	5.75	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.74	4.68	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.06	4.00	3.87	3.84	3.81	3.77	3.74	3.70	3.67
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.35	3.28	3.15	3.12	3.08	3.04	3.01	2.97	2.93
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.98	2.91	2.77	2.74	2.70	2.66	2.62	2.58	2.54
12	4.75	3.89	3.50	3.26	3.11	3.00	2.85	2.75	2.69	2.54	2.51	2.47	2.43	2.38	2.34	2.30
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.35	2.28	2.13	2.08	2.04	1.99	1.95	1.90	1.84
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.16	2.09	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.08	2.00	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.10	1.99	1.92	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.02	1.91	1.83	1.66	1.61	1.55	1.50	1.43	1.35	1.25
$\infty$	3.84	3.00	2.60	2.37	2.21	2.10	1.94	1.83	1.75	1.57	1.52	1.46	1.39	1.32	1.22	1.00

VALUES OF  $F_Q$  FOR  $Q(F | v_1, v_2) = 0.025$ 

$v_2 \backslash v_1$	1	2	3	4	5	6	8	10	12	20	24	30	40	60	120	$\infty$
1	647.8	799.5	864.2	899.6	921.8	937.1	956.7	968.6	976.7	993.1	997.2	1001	1006	1010	1014	1018
2	38.51	39.00	39.17	39.25	39.30	39.33	39.37	39.40	39.41	39.45	39.46	39.46	39.47	39.48	39.49	39.50
3	17.44	16.04	15.44	15.10	14.88	14.73	14.54	14.42	14.34	14.17	14.12	14.08	14.04	13.99	13.95	13.90
4	12.22	10.65	9.98	9.60	9.36	9.20	8.98	8.84	8.75	8.56	8.51	8.46	8.41	8.36	8.31	8.26
5	10.01	8.43	7.76	7.39	7.15	6.98	6.76	6.62	6.52	6.33	6.28	6.23	6.18	6.12	6.07	6.02
6	8.81	7.26	6.60	6.23	5.99	5.82	5.60	5.46	5.37	5.17	5.12	5.07	5.01	4.96	4.90	4.85
8	7.57	6.06	5.42	5.05	4.82	4.65	4.43	4.30	4.20	4.00	3.95	3.89	3.84	3.78	3.73	3.67
10	6.94	5.46	4.83	4.47	4.24	4.07	3.85	3.72	3.62	3.42	3.37	3.31	3.26	3.20	3.14	3.08
12	6.55	5.10	4.47	4.12	3.89	3.73	3.51	3.37	3.28	3.07	3.02	2.96	2.91	2.85	2.79	2.72
20	5.87	4.46	3.86	3.51	3.29	3.13	2.91	2.77	2.68	2.46	2.41	2.35	2.29	2.22	2.16	2.09
30	5.72	4.32	3.72	3.38	3.15	2.99	2.78	2.64	2.54	2.30	2.27	2.21	2.15	2.08	2.01	1.94
40	5.62	4.18	3.59	3.25	3.03	2.87	2.65	2.51	2.41	2.20	2.14	2.07	2.01	1.94	1.87	1.79
60	5.29	3.93	3.34	3.01	2.79	2.63	2.41	2.27	2.17	1.94	1.88	1.82	1.74	1.67	1.58	1.48
120	5.05	3.80	3.23	2.89	2.67	2.52	2.30	2.16	2.05	1.82	1.76	1.69	1.61	1.53	1.43	1.31
$\infty$	5.02	3.69	3.12	2.79	2.57	2.41	2.19	2.05	1.94	1.71	1.64	1.57	1.48	1.39	1.27	1.00

VALUES OF  $F_Q$  FOR  $Q(F | v_1, v_2) = 0.01$ 

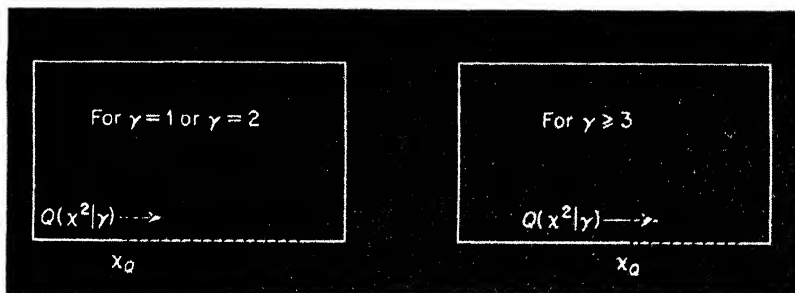
$v_1 \backslash v_2$	1	2	3	4	5	6	8	10	12	20	24	30	40	60	120	$\infty$
1	4052	5000	5403	5625	5764	5859	5982	6056	6106	6209	6235	6261	6287	6313	6339	6366
2	98.50	99.00	99.17	99.25	99.30	99.33	99.37	99.40	99.42	99.45	99.46	99.47	99.47	99.48	99.49	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.49	27.23	27.05	26.69	26.60	26.50	26.41	26.32	26.22	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.80	14.55	14.37	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.29	10.05	9.89	9.55	9.51	9.38	9.29	9.20	9.11	9.02
6	13.75	10.92	9.78	9.15	8.75	8.47	8.10	7.87	7.72	7.40	7.33	7.23	7.14	7.06	6.97	6.88
8	11.26	8.65	7.59	7.01	6.63	6.37	6.03	5.81	5.67	5.36	5.38	5.28	5.19	5.12	5.03	4.95
10	10.04	7.56	6.55	5.99	5.64	5.39	5.06	4.85	4.71	4.41	4.43	4.35	4.27	4.18	4.10	4.02
12	9.33	6.93	5.95	5.41	5.06	4.82	4.50	4.30	4.16	3.86	3.78	3.70	3.62	3.54	3.45	3.36
20	8.10	5.85	4.94	4.43	4.10	3.87	3.56	3.37	3.23	2.96	2.86	2.78	2.69	2.61	2.52	2.42
24	7.82	5.61	4.72	4.22	3.90	3.67	3.36	3.17	3.03	2.74	2.66	2.58	2.49	2.40	2.31	2.21
30	7.56	5.39	4.51	4.02	3.70	3.47	3.17	2.98	2.84	2.55	2.47	2.39	2.30	2.21	2.12	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	2.99	2.80	2.66	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.82	2.63	2.50	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.66	2.47	2.34	2.03	1.95	1.86	1.76	1.66	1.55	1.38
$\infty$	6.63	4.61	3.78	3.32	3.02	2.80	2.51	2.32	2.18	1.88	1.79	1.70	1.59	1.47	1.32	1.00

VALUES OF  $F_Q$  FOR  $Q(F | v_1, v_2) = 0.001$ 

$v_1 \backslash v_2$	1	2	3	4	5	6	8	10	12	20	24	30	40	60	120	$\infty$
1*	405.3	500.0	540.4	562.5	576.4	585.9	598.1	605.6	610.7	620.9	623.5	626.1	628.7	631.3	634.0	636.6
2	99.5	99.00	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.5	99.5	99.5	99.5	99.5	99.5
3	167.0	148.5	141.1	137.1	134.6	132.8	130.6	129.2	128.3	126.4	125.9	125.4	125.0	124.5	124.0	123.5
4	74.14	61.25	56.18	53.44	51.71	50.53	49.00	48.05	47.41	46.10	45.77	45.43	45.09	44.75	44.40	44.05
5	47.18	37.12	33.20	31.09	29.75	28.84	27.64	26.92	26.42	25.39	25.14	24.87	24.60	24.33	24.06	23.79
6	35.51	27.00	23.70	21.92	20.81	20.03	19.03	18.41	17.99	17.12	16.89	16.67	16.44	16.21	15.99	15.75
8	25.42	18.49	15.83	14.39	13.49	12.86	12.04	11.54	11.19	10.48	10.30	10.11	9.92	9.73	9.53	9.33
10	21.04	14.91	12.55	11.28	10.48	9.92	9.20	8.75	8.45	7.80	7.64	7.47	7.30	7.12	6.94	6.76
12	18.64	12.97	10.80	9.63	8.89	8.38	7.71	7.29	7.00	6.40	6.25	6.09	5.93	5.76	5.59	5.42
20	14.82	9.95	8.10	7.10	6.46	6.02	5.44	5.08	4.82	4.29	4.15	4.00	3.86	3.70	3.54	3.38
24	14.03	9.34	7.55	6.59	5.98	5.55	4.99	4.64	4.39	3.87	3.74	3.59	3.45	3.29	3.14	2.97
30	13.29	8.77	7.05	6.12	5.53	5.12	4.58	4.24	4.00	3.49	3.36	3.22	3.07	2.92	2.76	2.59
40	12.61	8.25	6.60	5.70	5.13	4.73	4.21	3.87	3.64	3.15	3.01	2.87	2.73	2.57	2.41	2.23
60	11.97	7.76	6.17	5.31	4.76	4.37	3.87	3.54	3.31	2.83	2.69	2.55	2.41	2.25	2.08	1.89
120	11.38	7.32	5.79	4.95	4.42	4.04	3.55	3.24	3.02	2.53	2.40	2.26	2.11	1.95	1.76	1.54
$\infty$	10.83	6.91	5.42	4.62	4.10	3.74	3.27	2.96	2.74	2.27	2.13	1.99	1.84	1.66	1.45	1.00

\* Multiply all entries on this line by 1000.

**APPENDIX 6: Values of  $\chi_Q^2$  for Selected Values of  $Q(\chi^2 | \nu)$**



For large values of  $\nu$ , ( $\nu > 30$ ).

$$\chi_Q^2 \doteq \left( 1 - \frac{2}{9\nu} \pm z_Q \sqrt{\frac{2}{9\nu}} \right)^3$$

where  $z_Q$  is the normal deviate cutting off the corresponding tails of a normal distribution.

For very large values of  $\nu$ , ( $\nu > 100$ ),

$$\chi_Q^2 \doteq \frac{1}{2}(z_Q \pm \sqrt{2\nu - 1})^2$$

This table is abridged from Table 8 of E. S. Pearson and H. O. Hartley, eds., *Biometrika Tables for Statisticians*, Vol. I (Cambridge: Cambridge University Press, 1954) and Table IV of R. A. Fisher and F. Yates, *Statistical Tables for Biological, Agricultural, and Medical Research*, 5th ed. (Edinburgh: Oliver and Boyd, Ltd., 1957), by permission.

VALUES OF  $\chi^2_{\alpha}$  FOR SELECTED VALUES OF  $Q(\chi^2 | \nu)$ 

$\nu$	0.999	0.995	0.99	0.98	0.975	0.95	0.90	0.80	0.75	0.70	0.50
1	0.0157	0.0433	0.0157	0.0628	0.0982	0.0393	0.0158	0.0642	0.102	0.148	0.455
2	0.00200	0.0100	0.0201	0.0404	0.0506	0.103	0.211	0.446	0.575	0.713	1.386
3	0.0243	0.0717	0.115	0.185	0.216	0.352	1.584	1.005	1.213	1.424	2.366
4	0.0908	0.207	0.297	0.429	0.484	0.711	1.649	1.649	1.923	2.195	3.357
5	0.210	0.412	0.554	0.752	0.831	1.145	1.610	2.343	2.675	3.000	4.351
6	0.381	0.676	0.872	1.134	1.237	1.635	2.204	3.070	3.455	3.828	5.348
7	0.598	0.989	1.239	1.564	1.690	2.167	2.833	3.822	4.255	4.671	6.346
8	0.857	1.344	1.646	2.032	2.180	2.733	3.490	4.594	5.071	5.527	7.344
9	1.152	1.735	2.088	2.532	2.700	3.325	4.168	5.380	5.899	6.343	8.343
10	1.479	2.156	2.558	3.059	3.247	3.940	4.865	6.179	6.737	7.267	9.342
11	1.834	2.603	3.053	3.609	3.816	4.575	5.578	6.989	7.584	8.148	10.341
12	2.214	3.074	3.571	4.178	4.404	5.226	6.304	7.807	8.438	9.034	11.340
13	2.617	3.565	4.107	4.765	5.009	5.892	7.042	8.634	9.299	9.926	12.340
14	3.041	4.075	4.660	5.368	5.629	6.571	7.790	9.467	10.165	10.821	13.339
15	3.483	4.601	5.229	5.985	6.262	7.261	8.547	10.307	11.036	11.721	14.339
16	3.942	5.142	5.812	6.614	6.908	7.962	9.312	11.152	11.912	12.624	15.338
17	4.416	5.697	6.408	7.255	7.564	8.672	10.085	12.002	12.792	13.531	16.338
18	4.905	6.265	7.015	7.906	8.231	9.390	10.865	12.857	13.675	14.440	17.338
19	5.407	6.844	7.633	8.567	8.907	10.117	11.651	13.716	14.562	15.352	18.338
20	5.921	7.434	8.260	9.237	9.591	10.851	12.443	14.578	15.452	16.266	19.337
21	6.447	8.034	8.897	9.915	10.283	11.591	13.240	15.445	16.344	17.182	20.337
22	6.983	8.643	9.542	10.600	10.982	12.338	14.041	16.314	17.240	18.101	21.337
23	7.529	9.260	10.196	11.293	11.688	13.091	14.848	17.187	18.137	19.021	22.337
24	8.085	9.886	10.856	11.992	12.401	13.848	15.659	18.062	19.037	19.943	23.337
25	8.649	10.520	11.524	12.697	13.120	14.611	16.473	18.940	19.930	20.867	24.337
26	9.222	11.160	12.198	13.409	13.844	15.379	17.292	19.820	20.843	21.792	25.336
27	9.803	11.808	12.879	14.125	14.573	16.151	18.114	20.703	21.749	22.719	26.336
28	10.391	12.461	13.565	14.847	15.308	16.928	18.939	21.588	22.657	23.647	27.336
29	10.986	13.121	14.256	15.574	16.047	17.708	19.768	22.475	23.567	24.577	28.336
30	11.588	13.787	14.953	16.306	16.791	18.493	20.599	23.364	24.478	25.508	29.336



VALUES OF  $\chi^2_0$  FOR SELECTED VALUES OF  $Q(\chi^2 | \nu)$ 

$\nu$	0.30	0.25	0.20	0.10	0.05	0.025	0.02	0.01	0.005	0.001
1	1.074	1.323	1.642	2.706	3.841	5.024	5.412	6.635	7.879	10.827
2	2.408	2.773	3.219	4.605	5.991	7.378	7.824	9.210	10.597	13.815
3	3.665	4.108	4.642	6.251	7.815	9.348	9.837	11.345	12.838	16.266
4	4.878	5.385	5.989	7.779	9.488	11.143	11.668	13.277	14.860	18.467
5	6.064	6.626	7.289	9.236	11.070	12.832	13.388	15.086	16.750	20.515
6	7.231	7.841	8.558	10.645	12.592	14.449	15.033	16.812	18.548	22.457
7	8.383	9.037	9.803	12.017	14.067	16.013	16.622	18.475	20.278	24.322
8	9.524	10.219	11.030	13.362	15.507	17.535	18.168	20.090	21.955	26.125
9	10.656	11.389	12.242	14.684	16.919	19.023	19.679	21.666	23.589	27.877
10	11.781	12.549	13.442	15.987	18.307	20.483	21.161	23.209	25.188	29.588
11	12.899	13.701	14.631	17.275	19.675	21.920	22.618	24.725	26.757	31.264
12	14.011	14.845	15.812	18.549	21.026	23.337	24.054	26.217	28.300	32.909
13	15.119	15.984	16.985	19.812	22.362	24.736	25.472	27.688	29.819	34.528
14	16.222	17.117	18.151	21.064	23.685	26.119	26.873	29.141	31.319	36.123
15	17.322	18.245	19.311	22.307	24.996	27.488	28.259	30.578	32.801	37.697
16	18.418	19.369	20.465	23.542	26.296	28.845	29.633	32.000	34.267	39.252
17	19.511	20.489	21.615	24.769	27.587	30.191	30.995	33.409	35.718	40.790
18	20.601	21.605	22.760	25.989	28.869	31.526	32.346	34.805	37.156	42.312
19	21.689	22.718	23.900	27.204	30.144	32.852	33.687	36.191	38.582	43.820
20	22.775	23.828	25.038	28.412	31.410	34.170	35.020	37.566	39.997	45.315
21	23.858	24.935	26.171	29.615	32.671	35.479	36.343	38.932	41.401	46.797
22	24.939	26.039	27.301	30.813	33.924	36.781	37.659	40.289	42.796	48.268
23	26.018	27.141	28.429	32.007	35.172	38.076	38.968	41.638	44.181	49.728
24	27.096	28.241	29.553	33.196	36.415	39.364	40.270	42.980	45.558	51.179
25	28.172	29.339	30.675	34.382	37.652	40.646	41.566	44.314	46.928	52.620
26	29.246	30.434	31.795	35.563	38.885	41.923	42.856	45.642	48.290	54.052
27	30.319	31.528	32.912	36.741	40.113	43.194	44.140	46.963	49.645	55.476
28	31.391	32.620	34.027	37.916	41.337	44.461	45.419	48.278	50.993	56.893
29	32.461	33.711	35.139	39.087	42.557	45.722	46.693	49.588	52.336	58.302
30	33.530	34.800	36.250	40.256	43.773	46.979	47.962	50.892	53.672	59.703

# **APPENDIX 7a: Number of Combinations of $N$ Things Taken $n$ at a Time: Binomial Coefficients**

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}$$

401

$\begin{matrix} N \\ n \end{matrix}$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2	1	1	3	6	10	15	21	28	36	45	55	66	78	91	105	120	136	153	171	190
3		1	1	4	10	20	35	56	84	120	165	220	286	364	455	560	680	816	969	1,140
4				1	5	15	35	70	126	210	330	495	715	1,001	1,365	1,820	2,380	3,060	3,876	4,845
5					1	6	21	56	126	252	462	792	1,287	2,002	3,003	4,368	6,188	8,568	11,628	15,504
6						1	7	28	84	210	462	924	1,716	3,003	5,005	8,008	12,376	18,564	27,132	38,760
7							1	8	36	120	330	792	1,716	3,432	6,435	11,440	19,448	31,824	50,888	77,520
8								1	9	45	165	495	1,287	3,003	6,435	12,870	24,310	43,758	75,582	125,970
9									1	10	55	220	715	2,002	5,005	11,440	24,310	48,620	92,378	167,960
10										1	11	66	286	1,001	3,003	8,008	19,448	43,758	92,378	184,756
11											1	12	78	364	1,365	4,368	12,376	31,824	75,582	167,960
12												1	13	91	455	1,820	6,188	18,564	50,888	125,970
13													1	14	105	560	2,380	8,568	27,132	77,520
14														1	15	120	680	3,060	11,628	38,760
15															1	16	136	816	3,876	15,504
16																1	17	153	969	4,845
17																	1	18	171	1,140
18																		1	19	190
19																			1	20
20																				1

Reprinted from Dudley J. Cowden, Statistical Methods in Quality Control (Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1957), pp. 652-653. Coefficients through  $\binom{100}{50}$  are published in Appendix III of Thornton C. Fry, Probability and its Engineering Uses (New York: D. Van Nostrand Co., Inc., 1928). These coefficients are rounded to 8 digits, except  $\binom{100}{50}$ , which is rounded to 9 digits.

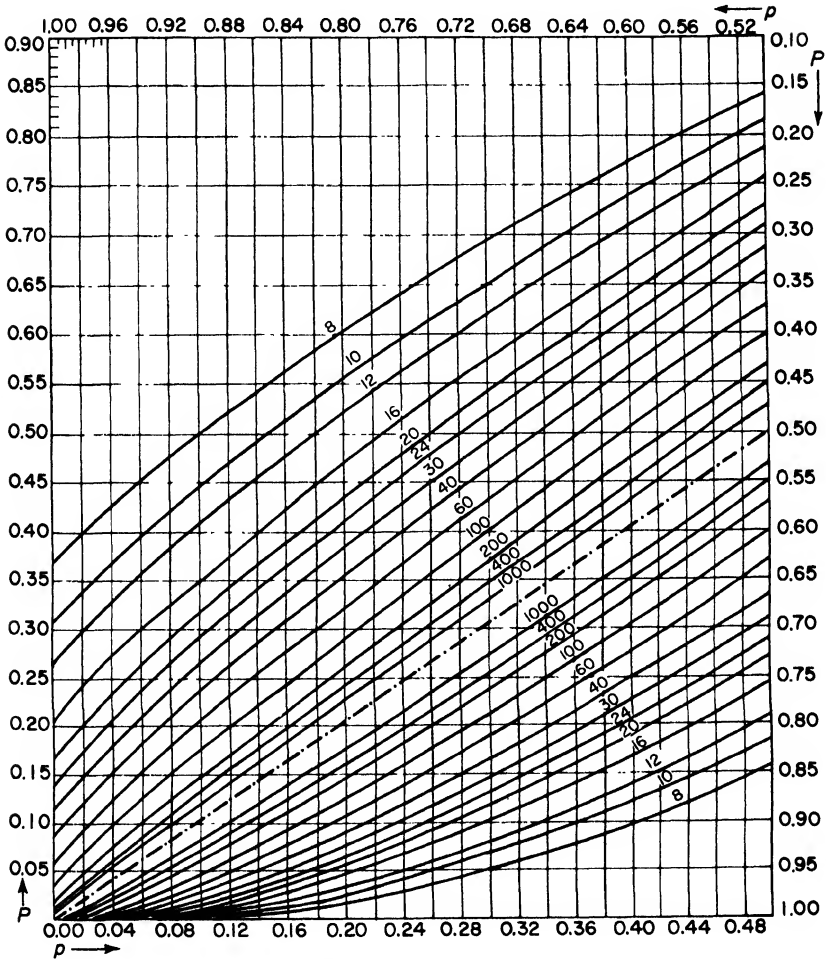
APPENDIX 7-b: Values of  $e^{-a}$  for Selected Values of  $a$

$a$	$e^{-a}$	$a$	$e^{-a}$	$a$	$e^{-a}$	$a$	$e^{-a}$
0.00	1.00000	0.25	0.77880	0.50	0.60653	0.75	0.47237
0.01	0.99005	0.26	0.77105	0.51	0.60050	0.76	0.46767
0.02	0.98020	0.27	0.76338	0.52	0.59452	0.77	0.46301
0.03	0.97045	0.28	0.75578	0.53	0.58860	0.78	0.45841
0.04	0.96079	0.29	0.74826	0.54	0.58275	0.79	0.45384
0.05	0.95123	0.30	0.74082	0.55	0.57695	0.80	0.44933
0.06	0.94176	0.31	0.73345	0.56	0.57121	0.81	0.44486
0.07	0.93239	0.32	0.72615	0.57	0.56553	0.82	0.44043
0.08	0.92312	0.33	0.71892	0.58	0.55990	0.83	0.43605
0.09	0.91393	0.34	0.71177	0.59	0.55433	0.84	0.43171
0.10	0.90484	0.35	0.70469	0.60	0.54881	0.85	0.42741
0.11	0.89583	0.36	0.69768	0.61	0.54335	0.86	0.42316
0.12	0.88692	0.37	0.69073	0.62	0.53794	0.87	0.41895
0.13	0.87810	0.38	0.68386	0.63	0.53259	0.88	0.41478
0.14	0.86936	0.39	0.67706	0.64	0.52729	0.89	0.41066
0.15	0.86071	0.40	0.67032	0.65	0.52205	0.90	0.40657
0.16	0.85214	0.41	0.66365	0.66	0.51685	0.91	0.40252
0.17	0.84366	0.42	0.65705	0.67	0.51171	0.92	0.39852
0.18	0.83527	0.43	0.65051	0.68	0.50662	0.93	0.39455
0.19	0.82696	0.44	0.64404	0.69	0.50158	0.94	0.39063
0.20	0.81873	0.45	0.63763	0.70	0.49659	0.95	0.38674
0.21	0.81058	0.46	0.63128	0.71	0.49164	0.96	0.38289
0.22	0.80252	0.47	0.62500	0.72	0.48675	0.97	0.37908
0.23	0.79453	0.48	0.61878	0.73	0.48191	0.98	0.37531
0.24	0.78663	0.49	0.61263	0.74	0.47711	0.99	0.37158

Note that  $e^{-(a+b)} = (e^{-a})(e^{-b})$ . Thus  $e^{-1.5} = (e^{-1.0})(e^{-0.5}) = 0.36788(0.60653) = 0.22313$ . Using this technique, one may use this table to find values of  $e^{-a}$  for selected values of  $a$  between 0.0 and 20.0 with 5 decimal place accuracy.

**APPENDIX 8: Charts for Obtaining Confidence Limits for  $P$  in Binomial Sampling, Given  $d$  and  $n$ ;  
( $p = d/n$ )**

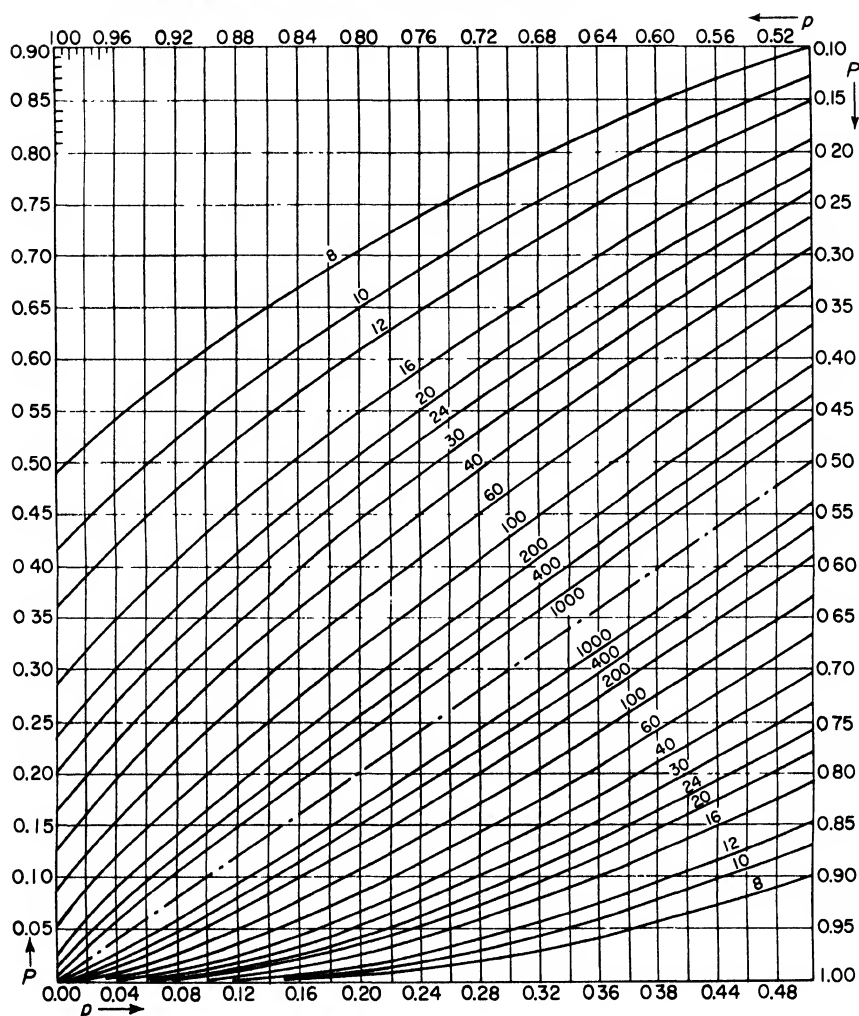
**CHART A 8.1: 95 PER CENT CONFIDENCE LIMITS**



*The confidence is always equal or less than that stated.*

The numbers printed along the curves indicate the sample size  $n$ . If for a given value of the abscissa  $p$ ,  $P_1$  and  $P_2$  are the ordinates read from (or interpolated between) the appropriate lower and upper curves, then  $\text{Prob}(P_1 < P < P_2) < 1 - \alpha$ .

CHART A 8.2: 99 PER CENT CONFIDENCE LIMITS



*The numbers printed along the curves indicate the sample size of  $n$ . Note: The process of reading from the curves can be simplified with the help of the right-angled corner of a loose sheet of paper or thin card, along the edges of which are marked off the scales shown in the top left-hand corner of each chart. Taken, by permission, from E. S. Pearson and H. O. Hartley, eds., *Biometrika Tables for Statisticians*, Vol. I (Cambridge: Cambridge University Press, 1954), Table 41, pp. 204-205.*

**APPENDIX 9: Factors for Obtaining Unbiased Estimates of  $\sigma$ , and for Obtaining Control Limits for Ranges, when Sampling from a Normal Population**

$$E(a_0R) = \sigma \quad E[a_1(SD)] = \sigma \quad E(a_2s) = \sigma \quad LCL(R) = D_3\bar{R} \quad UCL(R) = D_4\bar{R}$$

Sample size	$a_0$	$a_1$	$a_2$	$D_3$	$D_4$
2	0.8865	1.7725	1.2533	...	3.2665
3	0.5910	1.3820	1.1284	...	2.5746
4	0.4859	1.2533	1.0854	...	2.2820
5	0.4299	1.1894	1.0638	...	2.1145
6	0.3946	1.1512	1.0509	...	2.0038
7	0.3698	1.1259	1.0424	0.0757	1.9243
8	0.3511	1.1078	1.0362	0.1362	1.8638
9	0.3367	1.0942	1.0317	0.1840	1.8160
10	0.3249	1.0837	1.0281	0.2230	1.7770
11	0.3153	1.0753	1.0253	0.2556	1.7444
12	0.3069	1.0684	1.0230	0.2833	1.7167
13	0.2998	1.0627	1.0210	0.3072	1.6928
14	0.2936	1.0579	1.0194	0.3281	1.6719
15	0.2880	1.0537	1.0180	0.3466	1.6534
16	0.2831	1.0501	1.0168	0.3630	1.6370
17	0.2787	1.0470	1.0157	0.3778	1.6222
18	0.2747	1.0442	1.0148	0.3913	1.6087
19	0.2711	1.0418	1.0140	0.4035	1.5965
20	0.2678	1.0396	1.0132	0.4147	1.5853
21	0.265	1.0376	1.0126	...	...
22	0.262	1.0358	1.0120	...	...
23	0.259	1.0342	1.0114	...	...
24	0.256	1.0327	1.0109	...	...
25	0.254	1.0313	1.0105	...	...

$$a_0 = \sigma/E(R); a_1 = g\sqrt{n/2}; a_2 = a_1\sqrt{(n-1)/n}; g = \Gamma\left(\frac{n-1}{2}\right)/\Gamma\left(\frac{n}{2}\right); \Gamma(0.5) =$$

$$\sqrt{\pi}; \Gamma(n+1) = n!; D_3 = \frac{E(R) - 3\sigma_R}{E(R)}; D_4 = \frac{E(R) + 3\sigma_R}{E(R)}$$

Entries for  $a_0$ ,  $D_3$ , and  $D_4$  were calculated by using values of the expected value and standard deviation of the range as given in: E. S. Pearson and H. O. Hartley, eds., *Biometrika Tables for Statisticians, Vol. I* (Cambridge: Cambridge University Press, 1966), p. 176. There can be no lower control limit for ranges when  $n < 7$ , because  $E(R) - 3\sigma_R < 0$ . For a more extensive table of values of  $a_1$  and  $a_2$  see Ben W. Bolch "More on Unbiased Estimation of the Standard Deviation," *American Statistician*, June, 1968, p. 27.

APPENDIX 10: Values of  $r_0$  for Selected Values of  $Q(r|v)$  when  $\rho = 0$ 

$v$	0.45	0.40	0.35	0.30	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005
1	0.156	0.309	0.454	0.588	0.707	0.809	0.891	0.951	0.988	0.997	0.999*	0.999*
2	0.100	0.200	0.300	0.400	0.500	0.600	0.700	0.800	0.900	0.950	0.980	0.990
3	0.079	0.158	0.238	0.320	0.404	0.492	0.585	0.680	0.805	0.878	0.934	0.959
4	0.067	0.134	0.203	0.274	0.347	0.426	0.511	0.608	0.729	0.811	0.882	0.917
5	0.059	0.119	0.180	0.243	0.309	0.380	0.459	0.551	0.669	0.755	0.833	0.875
6	0.053	0.108	0.163	0.220	0.281	0.347	0.420	0.507	0.621	0.707	0.789	0.834
7	0.049	0.099	0.150	0.203	0.260	0.321	0.390	0.472	0.582	0.666	0.750	0.798
8	0.046	0.092	0.140	0.190	0.242	0.300	0.365	0.443	0.549	0.632	0.715	0.765
9	0.043	0.087	0.132	0.178	0.228	0.282	0.344	0.419	0.521	0.602	0.685	0.735
10	0.041	0.082	0.125	0.169	0.216	0.268	0.327	0.398	0.497	0.576	0.658	0.708
11	0.039	0.078	0.119	0.161	0.206	0.255	0.312	0.380	0.476	0.553	0.634	0.684
12	0.037	0.075	0.113	0.154	0.197	0.244	0.298	0.365	0.457	0.532	0.612	0.661
13	0.035	0.072	0.109	0.148	0.189	0.235	0.287	0.351	0.441	0.514	0.592	0.641
14	0.034	0.069	0.104	0.142	0.182	0.226	0.276	0.338	0.426	0.497	0.574	0.623
15	0.033	0.066	0.101	0.137	0.176	0.218	0.267	0.327	0.412	0.482	0.558	0.606
16	0.032	0.064	0.098	0.133	0.170	0.211	0.259	0.317	0.400	0.468	0.542	0.590
17	0.031	0.062	0.095	0.128	0.165	0.205	0.251	0.308	0.389	0.456	0.529	0.575
18	0.030	0.060	0.092	0.125	0.160	0.199	0.244	0.299	0.378	0.444	0.515	0.561
19	0.029	0.059	0.089	0.121	0.156	0.194	0.238	0.291	0.369	0.433	0.503	0.549
20	0.028	0.057	0.087	0.118	0.152	0.189	0.231	0.284	0.360	0.423	0.492	0.537
21	0.028	0.056	0.085	0.115	0.148	0.184	0.226	0.277	0.352	0.413	0.482	0.526
22	0.027	0.054	0.083	0.113	0.145	0.180	0.221	0.271	0.344	0.404	0.472	0.515
23	0.026	0.053	0.081	0.110	0.141	0.176	0.216	0.265	0.337	0.396	0.462	0.505
24	0.026	0.052	0.079	0.108	0.138	0.172	0.211	0.260	0.330	0.388	0.453	0.496
25	0.025	0.051	0.078	0.106	0.136	0.169	0.207	0.255	0.323	0.381	0.445	0.487
26	0.025	0.050	0.076	0.104	0.133	0.166	0.203	0.250	0.317	0.374	0.437	0.479
27	0.024	0.049	0.075	0.102	0.131	0.162	0.199	0.245	0.311	0.367	0.430	0.471
28	0.024	0.048	0.073	0.100	0.128	0.160	0.196	0.241	0.306	0.361	0.423	0.463
29	0.024	0.047	0.072	0.098	0.126	0.157	0.192	0.237	0.301	0.355	0.416	0.456
30	0.023	0.047	0.071	0.096	0.124	0.154	0.189	0.233	0.296	0.349	0.409	0.449
40	0.020	0.040	0.061	0.083	0.107	0.133	0.164	0.202	0.257	0.304	0.358	0.393
60	0.016	0.033	0.050	0.068	0.087	0.109	0.134	0.165	0.211	0.250	0.295	0.325
120	0.012	0.023	0.035	0.048	0.062	0.077	0.095	0.117	0.150	0.178	0.210	0.232

\* Greater than 0.999.

For simple correlation  $v = n - 2$ . For partial correlation  $v = n - k - 2$ , where  $k$  is the number of variables held constant. Table

APPENDIX 11: Values of  $z_r$  for Selected Values of  $r$ 

$r$	$z_r$	$r$	$z_r$	$r$	$z_r$	$r$	$z_r$
0.00	0.00000	0.25	0.25541	0.50	0.54931	0.75	0.97296
0.01	0.01000	0.26	0.26611	0.51	0.56273	0.76	0.99622
0.02	0.02000	0.27	0.27686	0.52	0.57634	0.77	1.02033
0.03	0.03001	0.28	0.28768	0.53	0.59015	0.78	1.04537
0.04	0.04002	0.29	0.29857	0.54	0.60416	0.79	1.07143
0.05	0.05004	0.30	0.30952	0.55	0.61838	0.80	1.09861
0.06	0.06007	0.31	0.32055	0.56	0.63283	0.81	1.12703
0.07	0.07011	0.32	0.33165	0.57	0.64752	0.82	1.15682
0.08	0.08017	0.33	0.34283	0.58	0.66246	0.83	1.18814
0.09	0.09024	0.34	0.35409	0.59	0.67767	0.84	1.22117
0.10	0.10034	0.35	0.36544	0.60	0.69315	0.85	1.25615
0.11	0.11045	0.36	0.37689	0.61	0.70892	0.86	1.29334
0.12	0.12058	0.37	0.38842	0.62	0.72501	0.87	1.33308
0.13	0.13074	0.38	0.40006	0.63	0.74142	0.88	1.37577
0.14	0.14093	0.39	0.41180	0.64	0.75817	0.89	1.42193
0.15	0.15114	0.40	0.42365	0.65	0.77530	0.90	1.47222
0.16	0.16139	0.41	0.43561	0.66	0.79281	0.91	1.52752
0.17	0.17167	0.42	0.44769	0.67	0.81074	0.92	1.58903
0.18	0.18198	0.43	0.45990	0.68	0.82911	0.93	1.65839
0.19	0.19234	0.44	0.47223	0.69	0.84796	0.94	1.73805
0.20	0.20273	0.45	0.48470	0.70	0.86730	0.95	1.83178
0.21	0.21317	0.46	0.49731	0.71	0.88718	0.96	1.94591
0.22	0.22366	0.47	0.51007	0.72	0.90764	0.97	2.09230
0.23	0.23419	0.48	0.52298	0.73	0.92873	0.98	2.29756
0.24	0.24477	0.49	0.53606	0.74	0.95048	0.99	2.64665

$$z_r = \operatorname{arctanh} r = \frac{1}{2} \log_e \left( \frac{1+r}{1-r} \right).$$



## APPENDIX 12: Random Numbers

1581922396	2068577984	8262130892	8374856049	4637567488
0928105582	7295088579	9586111652	7055508767	6472382934
4112077556	3440672486	1882412963	0684012006	0933147914
7457477468	5435810788	9670852913	1291265730	4890031305
0099520858	3090908872	2039593181	5973470495	9776135501
7245174840	2275698645	8416549348	4676463101	2229367983
6749420382	4832630032	5670984959	5432114610	2966095680
5503161011	7413686599	1198757695	0414294470	0140121598
7164238934	7666127259	5263097712	5133648980	4011966963
3593969525	0272759769	0385998136	9999089966	7544056852
4192054466	0700014629	5169439659	8408705169	1074373131
9697426117	6488888550	4031652526	8123543276	0927534537
2007950579	9564268448	3457416988	1531027886	7016633739
4584768758	2389278610	3859431781	3643768456	4141314518
3840145867	9120831830	7228567652	1267173884	4020651657
0190453442	4800088084	1165628559	5407921254	3768932478
6766554338	5585265145	5089052204	9780623691	2195448096
6315116284	9172824179	5544814339	0016943666	3828538786
3908771938	4035554324	0840126299	4942059208	1475623997
5570024586	9324732596	1186563397	4425143189	3216653251
2999997185	0135968938	7678931194	1351031403	6002561840
7864375912	8383232768	1892857070	2323673751	3188881718
7065492027	6349104233	3382569662	4579426926	1513082455
0654683246	4765104877	8149224168	5468631609	6474393896
7830555058	5255147182	3519287786	2481675649	8907598697
7626984369	4725370390	9641916289	5049082870	7463807244
4785048453	3646121751	8436077768	2928794356	9956043516
4627791048	5765558107	8762592043	6185670830	6363845920
9376470693	0441608934	8749472723	2202271078	5897002653
1227991661	7936797054	9527542791	4711871173	8300978148
5582095589	5535798279	4764439855	6279247618	4446895088
4959397698	1056981450	8416606706	8234013222	6426813469
1824779358	1333750468	9434074212	5273692238	5902177065
7041092295	5726289716	3420847871	1820481234	0318831723
3555104281	0903099163	6827824899	6383872737	5901682626
9717595534	1634107293	8521057472	1471300754	3044151557
5571564123	7344613447	1129117244	3208461091	1699403490
4674262892	2809456764	5806554509	8224980942	5738031833
8461228715	0746980892	9285305274	6331989646	8764467686
1838538678	3049068967	6955157269	5482964330	2161984904
1834182305	6203476893	5937802079	3445280195	3694915658
1884227732	2923727501	8044389132	4611203081	6072112445
6791857341	6696243386	2219599137	3193884236	8224729718
3007929946	4031562749	5570757297	6273785046	1455349704
6085440624	2875556938	5496629750	4841817356	1443167141
7005051056	3496332071	5054070890	7303867953	6255181190
9846413446	8306646692	0661684251	8875127201	6251533454
0625457703	4229164694	7321363715	7051128285	1108468072
5457593922	9751489574	1799906380	1989141062	5595364247
4076486653	8950826528	4934582003	4071187742	1456207629

## APPENDIX 13: Common Logarithms

Every logarithm has an integer part, called the *characteristic*, and a decimal part, called the *mantissa*. This table contains only mantissas. When  $X > 1$ , the characteristic is positive and is numerically one less than the number of places to the left of the decimal point. When  $X < 1$ , the characteristic is negative and is numerically one more than the number of zeros immediately to the right of the decimal point. Thus, the logarithm of 24.9 is 1.39620, and the logarithm of 0.0249 is  $\bar{2}.39620 - 10$ . The logarithm of any number less than or equal to zero is undefined.

The following rules illustrate computation with logarithms.

Rule	Example
1. $ab = \text{antilog}(\log a + \log b)$	$2(3)$ $= \text{antilog}(0.30103 + 0.47712)$ $= \text{antilog}(0.77815) = 6$
2. $b/a = \text{antilog}(\log b - \log a)$	$\frac{3}{2}$ $= \text{antilog}(0.47712 - 0.30103)$ $= \text{antilog}(0.17609) = 1.5$
3. $a^b = \text{antilog}[b(\log a)]$	$2^3$ $= \text{antilog}[3(0.30103)]$ $= \text{antilog}(0.90309) = 8$
4. $\sqrt[b]{a} = \text{antilog}[(1/b)(\log a)]$	$\sqrt[3]{2} = 2^{1/3}$ $= \text{antilog}[(\frac{1}{3})(0.30103)]$ $= 1.2599$

## Common Logarithms (cont.)

X	0	1	2	3	4	5	6	7	8	9	D
0	— ∞	00000	30103	47712	60206	69897	77815	84510	90309	95424	...
10	00000	00432	00860	01284	01703	02119	02531	02938	03342	03743	*
11	04139	04532	04922	05308	05690	06070	06446	06819	07188	07555	*
12	07918	08279	08636	08991	09342	09691	10037	10380	10721	11059	*
13	11394	11737	12057	12385	12710	13038	13364	13672	13988	14301	*
14	14613	14922	15229	15534	15836	16137	16435	16732	17026	17319	*
15	17609	17898	18184	18469	18752	19033	19312	19590	19866	20140	*
16	20412	20683	20952	21219	21484	21748	22011	22272	22531	22789	*
17	23045	23300	23553	23805	24055	24304	24551	24797	25042	25285	*
18	25527	25768	26007	26245	26482	26717	26951	27184	27416	27646	*
19	27875	28103	28330	28556	28780	29003	29226	29447	29667	29885	*
20	30103	30320	30535	30750	30963	31175	31387	31597	31806	32015	212
21	32222	32428	32634	32838	33041	33244	33445	33646	33846	34044	202
22	34242	34439	34635	34830	35025	35218	35411	35603	35793	35984	193
23	36173	36361	36549	36736	36922	37107	37291	37475	37658	37840	185
24	38021	38202	38382	38561	38739	38917	39094	39270	39445	39620	177
25	39794	39967	40140	40312	40483	40654	40824	40993	41162	41330	170
26	41497	41664	41830	41996	42160	42325	42488	42651	42813	42975	164
27	43136	43297	43457	43616	43775	43933	44091	44248	44404	44560	158
28	44716	44871	45025	45179	45332	45484	45637	45788	45939	46090	152
29	46240	46389	46538	46687	46835	46982	47129	47276	47422	47567	147
30	47712	47857	48001	48144	48287	48430	48572	48714	48855	48996	142
31	49136	49276	49415	49554	49693	49831	49969	50106	50243	50379	138
32	50515	50651	50786	50920	51055	51188	51322	51455	51587	51720	134
33	51851	51983	52114	52244	52375	52504	52634	52763	52892	53020	130
34	53148	53275	53403	53529	53656	53782	53908	54033	54158	54283	126
35	54407	54531	54654	54777	54900	55023	55145	55267	55388	55509	122
36	55630	55751	55871	55991	56110	56229	56348	56467	56585	56703	119
37	56820	56937	57054	57171	57287	57403	57519	57634	57749	57864	116
38	57978	58092	58206	58320	58433	58546	58659	58771	58883	58995	113
39	59106	59218	59329	59439	59550	59660	59770	59879	59988	60097	110
40	60206	60314	60423	60531	60638	60746	60853	60959	61066	61172	107
41	61278	61384	61490	61595	61700	61805	61909	62014	62118	62221	105
42	62325	62428	62531	62634	62737	62839	62941	63043	63144	63246	102
43	63347	63448	63548	63649	63749	63849	63949	64048	64147	64246	100
44	64345	64444	64542	64640	64738	64836	64933	65031	65128	65225	98
45	65321	65418	65514	65610	65706	65801	65896	65992	66087	66181	96
46	66276	66370	66464	66558	66652	66745	66839	66932	67025	67117	93
47	67210	67302	67394	67486	67578	67669	67761	67852	67943	68034	91
48	68124	68215	68305	68395	68485	68574	68664	68753	68842	68931	90
49	69020	69108	69197	69285	69373	69461	69548	69636	69723	69810	88

\* Logarithms of numbers from 1001 through 1999 may be obtained, without interpolation, on pages 412-413.

## Common Logarithms (cont.)

X	0	1	2	3	4	5	6	7	8	9	D
50	69897	69984	70070	70157	70243	70329	70415	70501	70586	70672	86
51	70757	70842	70927	71012	71096	71181	71265	71349	71433	71517	84
52	71600	71684	71767	71850	71933	72016	72099	72181	72263	72346	83
53	72428	72509	72591	72673	72754	72835	72916	72997	73078	73159	81
54	73239	73320	73400	73480	73560	73640	73719	73799	73878	73957	80
55	74036	74115	74194	74273	74351	74429	74507	74586	74663	74741	78
56	74819	74896	74974	75051	75128	75205	75282	75358	75435	75511	77
57	75587	75664	75740	75815	75891	75967	76042	76118	76193	76268	76
58	76343	76418	76492	76567	76641	76716	76790	76864	76938	77012	74
59	77085	77159	77232	77305	77379	77452	77525	77597	77670	77743	73
60	77815	77887	77960	78032	78104	78176	78247	78319	78390	78462	72
61	78533	78604	78675	78746	78817	78888	78958	79029	79099	79169	71
62	79239	79309	79379	79449	79518	79588	79657	79727	79796	79865	70
63	79934	80003	80072	80140	80209	80277	80346	80414	80482	80550	68
64	80618	80686	80754	80821	80889	80956	81023	81090	81158	81224	67
65	81291	81358	81425	81491	81558	81624	81690	81757	81823	81889	66
66	81954	82020	82086	82151	82217	82282	82347	82413	82478	82543	65
67	82607	82672	82737	82802	82866	82930	82995	83059	83123	83187	64
68	83251	83315	83378	83442	83506	83569	83632	83696	83759	83822	63
69	83885	83948	84011	84073	84136	84198	84261	84323	84386	84448	62
70	84510	84572	84634	84696	84757	84819	84880	84942	85003	85065	62
71	85126	85187	85248	85309	85370	85431	85491	85552	85612	85673	61
72	85733	85794	85854	85914	85974	86034	86094	86153	86213	86273	60
73	86332	86392	86451	86510	86570	86629	86688	86747	86806	86864	59
74	86923	86982	87040	87099	87157	87216	87274	87332	87390	87448	58
75	87506	87564	87622	87679	87737	87795	87852	87910	87967	88024	58
76	88081	88138	88195	88252	88309	88366	88423	88480	88536	88593	57
77	88649	88705	88762	88818	88874	88930	88986	89042	89098	89154	56
78	89209	89265	89321	89376	89432	89487	89542	89597	89653	89708	55
79	89763	89818	89873	89927	89982	90037	90091	90146	90200	90255	55
80	90309	90363	90417	90472	90526	90580	90634	90687	90741	90795	54
81	90849	90902	90956	91009	91062	91116	91169	91222	91275	91328	53
82	91381	91434	91487	91540	91593	91645	91698	91751	91803	91855	53
83	91908	91960	92012	92065	92117	92169	92221	92273	92324	92376	52
84	92428	92480	92531	92583	92634	92686	92737	92788	92840	92891	51
85	92942	92993	93044	93095	93146	93197	93247	93298	93349	93399	51
86	93450	93500	93551	93601	93651	93702	93752	93802	93852	93902	50
87	93952	94002	94052	94101	94151	94201	94250	94300	94349	94399	50
88	94448	94498	94547	94596	94645	94694	94743	94792	94841	94890	49
89	94939	94988	95036	95085	95134	95182	95231	95279	95328	95376	49
90	95424	95472	95521	95569	95617	95665	95713	95761	95809	95856	48
91	95904	95952	95999	96047	96095	96142	96190	96237	96284	96332	48
92	96379	96426	96473	96520	96567	96614	96661	96708	96755	96802	47
93	96848	96895	96942	96988	97035	97081	97128	97174	97220	97267	47
94	97313	97359	97405	97451	97497	97543	97589	97635	97681	97727	46
95	97772	97818	97864	97909	97955	98000	98046	98091	98137	98182	46
96	98227	98272	98318	98363	98408	98453	98498	98543	98588	98632	45
97	98677	98722	98767	98811	98856	98900	98945	98989	99034	99078	45
98	99123	99167	99211	99255	99300	99344	99388	99432	99476	99520	44
99	99564	99607	99651	99695	99739	99782	99826	99870	99913	99957	44

## Common Logarithms (cont.)

X	0	1	2	3	4	5	6	7	8	9	D
100	000000	000434	000868	001301	001734	002166	002598	003029	003461	003891	432
101	4321	4751	5181	5609	6038	6466	6894	7321	7748	8174	428
102	8600	9026	9451	9876	010300	010724	011147	011570	011993	012415	424
103	012837	013259	013680	014100	4521	4940	5360	5779	6197	6616	420
104	7033	7451	7868	8284	8700	9116	9532	9947	020361	020775	416
105	021189	021603	022016	022428	022841	023252	023664	024075	4486	4896	412
106	5306	5715	6125	6533	6942	7350	7757	8164	8571	8978	408
107	9384	9789	030195	030600	031004	031408	031812	032216	032619	033021	404
108	033424	033826	4227	4628	5029	5430	5830	6230	6629	7028	400
109	7426	7825	8223	8620	9017	9414	9811	040207	040602	040998	397
110	041893	041787	042182	042576	042969	043362	043755	044148	044540	044932	393
111	5323	5714	6105	6495	6885	7275	7664	8053	8442	8830	390
112	9218	9606	9993	050380	050766	051153	051538	051924	052309	052694	386
113	053078	053463	053846	4230	4613	4996	5378	5760	6142	6524	383
114	6905	7286	7666	8046	8426	8805	9185	9563	9942	060320	379
115	060698	061075	061452	061829	062206	062582	062958	063333	063709	4083	376
116	4458	4832	5206	5580	5953	6326	6699	7071	7443	7815	373
117	8186	8557	8928	9298	9668	070038	070407	070776	071145	071514	370
118	071882	072250	072617	072985	073352	3718	4085	4451	4816	5182	366
119	5547	5912	6276	6640	7004	7368	7731	8094	8457	8819	363
120	079181	079543	079904	080266	080626	080987	081347	081707	082067	082426	360
121	082785	083144	083503	3861	4219	4576	4934	5291	5647	6004	357
122	6360	6716	7071	7426	7781	8136	8490	8845	9198	9552	355
123	9905	090258	090611	090963	091315	091667	092018	092370	092721	093071	352
124	093422	3772	4122	4471	4820	5169	5518	5866	6215	6562	349
125	6910	7257	7604	7951	8298	8644	8990	9335	9681	100026	346
126	100371	100715	101059	101403	101747	102091	102434	102777	103119	3462	343
127	3804	4146	4487	4828	5169	5510	5851	6191	6531	6871	341
128	7210	7549	7888	8227	8565	8903	9241	9579	9916	110253	338
129	110590	110926	111263	111599	111934	112270	112605	112940	113275	3609	335
130	113943	114277	114611	114944	115278	115611	115943	116276	116608	116940	333
131	7271	7603	7934	8265	8595	8926	9256	9586	9915	120245	330
132	120574	120903	121231	121560	121888	122216	122544	122871	123198	3525	328
133	3852	4178	4504	4830	5156	5481	5806	6131	6456	6781	325
134	7105	7429	7753	8076	8399	8722	9045	9368	9690	130012	323
135	130334	130655	130977	131298	131619	131939	132260	132580	132900	3219	321
136	3539	3858	4177	4496	4814	5133	5451	5769	6086	6403	318
137	6721	7037	7354	7671	7987	8303	8618	8934	9249	9564	316
138	9879	140194	140508	140822	141136	141450	141763	142076	142389	142702	314
139	143015	3327	3639	3951	4263	4574	4885	5196	5507	5818	311
140	146128	146438	146748	147058	147367	147676	147985	148294	148603	148911	309
141	9219	9527	9835	150142	150449	150756	151063	151370	151676	151982	307
142	152288	152594	152900	3205	3510	3815	4120	4424	4728	5032	305
143	5336	5640	5943	6246	6549	6852	7154	7457	7759	8061	303
144	8362	8664	8965	9266	9567	9868	160168	160469	160769	161068	301
145	161368	161667	161967	162266	162564	162863	3161	3460	3758	4055	299
146	4353	4650	4947	5244	5541	5838	6134	6430	6726	7022	297
147	7317	7613	7908	8203	8497	8792	9086	9380	9674	9968	295
148	170262	170555	170848	171141	171434	171726	172019	172311	172603	172895	293
149	3186	3478	3769	4060	4351	4641	4932	5222	5512	5802	291

## Common Logarithms (cont.)

X	0	1	2	3	4	5	6	7	8	9	D
150	176091	176381	176670	176959	177248	177536	177825	178113	178401	178689	289
151	8977	9264	9552	9839	180126	180413	180699	180986	181272	181558	287
152	181844	182129	182415	182700	2985	3270	3555	3839	4123	4407	285
153	4691	4975	5259	5542	5825	6108	6391	6674	6956	7239	283
154	7521	7803	8084	8366	8647	8928	9209	9490	9771	190051	281
155	190332	190612	190892	191171	191451	191730	192010	192289	192567	2846	279
156	3125	3403	3681	3959	4237	4514	4792	5069	5346	5623	278
157	5900	6176	6453	6729	7005	7281	7556	7832	8107	8382	276
158	8657	8932	9206	9481	9755	200029	200303	200577	200850	201124	274
159	201397	201670	201943	202216	202488	2761	3033	3305	3577	3848	272
160	204120	204391	204663	204934	205204	205475	205746	206016	206286	206556	271
161	6826	7096	7365	7634	7904	8173	8441	8710	8979	9247	269
162	9515	9783	210051	210319	210586	210853	211121	211388	211654	211921	267
163	212188	212454	2720	2986	3252	3518	3783	4049	4314	4579	266
164	4844	5109	5373	5638	5902	6166	6430	6694	6957	7221	264
165	7484	7747	8010	8273	8536	8798	9060	9323	9585	9846	262
166	220108	220370	220631	220892	221153	221414	221675	221936	222196	222456	261
167	2718	2976	3238	3496	3755	4015	4274	4533	4792	5051	259
168	5309	5568	5826	6084	6342	6600	6858	7115	7372	7630	258
169	7887	8144	8400	8657	8913	9170	9426	9682	9938	230193	256
170	230449	230704	230960	231215	231470	231724	231979	232234	232488	232742	255
171	2996	3250	3504	3757	4011	4264	4517	4770	5023	5276	253
172	5528	5781	6033	6285	6537	6789	7041	7292	7544	7795	252
173	8046	8297	8548	8799	9049	9299	9550	9800	240050	240300	250
174	240549	240799	241048	241297	241546	241795	242044	242293	2541	2790	249
175	3038	3286	3534	3782	4030	4277	4525	4772	5019	5266	248
176	5513	5759	6006	6252	6499	6745	6991	7237	7482	7728	246
177	7973	8219	8464	8709	8954	9198	9443	9687	9932	250176	245
178	250420	250664	250908	251151	251395	251638	251881	252125	252368	2610	243
179	2853	3096	3338	3580	3822	4064	4306	4548	4790	5031	242
180	255273	255514	255755	255996	256237	256477	256718	256958	257198	257439	241
181	7679	7918	8158	8398	8637	8877	9116	9355	9594	9833	239
182	260071	260310	260548	260787	261025	261263	261501	261739	261976	262214	238
183	2451	2688	2925	3162	3399	3636	3873	4109	4346	4582	237
184	4818	5054	5290	5525	5761	5996	6232	6467	6702	6937	235
185	7172	7406	7641	7875	8110	8344	8578	8812	9046	9279	234
186	9513	9746	9980	270213	270446	270679	270912	271144	271377	271609	233
187	271842	272074	272306	2538	2770	3001	3233	3464	3696	3927	232
188	4158	4389	4620	4850	5081	5311	5542	5772	6002	6232	230
189	6462	6692	6921	7151	7380	7609	7838	8067	8296	8525	229
190	278754	278982	279211	279439	279667	279895	280123	280351	280578	280806	228
191	281033	281261	281488	281715	281942	282169	282396	282622	282849	3075	227
192	3301	3527	3753	3979	4205	4431	4656	4882	5107	5332	226
193	5557	5782	6007	6232	6456	6681	6905	7130	7354	7578	225
194	7802	8026	8249	8473	8696	8920	9143	9366	9589	9812	223
195	290035	290257	290480	290702	290925	291147	291369	291591	291813	292034	222
196	2256	2478	2699	2920	3141	3363	3584	3804	4025	4246	221
197	4466	4687	4907	5127	5347	5567	5787	6007	6226	6446	220
198	6665	6884	7104	7323	7542	7761	7979	8198	8416	8635	219
199	8853	9071	9289	9507	9725	9943	300161	300378	300595	300813	218

---

### APPENDIX 14: Squares, Square Roots, and Reciprocals

Although these tables are for values of  $n$  with three or fewer digits, they can be used for values of  $n$  containing more than three significant digits with an answer significant to five digits (usually), by the following simple device.

1. Let  $n'$  be the value of  $n$  recorded in Appendix 14-A or 14-B closest to the desired value of  $n$ . If  $n$  has an odd number of digits to the left of the decimal point, or is less than one with an odd number of zeros to the right of the decimal point, e.g., 0.023, use Appendix 14-A. If  $n$  has an even number of digits to the left of the decimal point, or is less than one with an even number of zeros to the right of the decimal point, e.g., 0.0023, use Appendix 14-B.

2. Look up  $\sqrt{n'}$  in Appendix 14-A or 14-B.

3. Compute  $\sqrt{n} \doteq \frac{n + n'}{2\sqrt{n'}}$ .

## Squares, Square Roots, and Reciprocals (cont.)

$n$	$n^2$	$\sqrt{n}$	$1/n$	$n$	$n^2$	$\sqrt{n}$	$1/n$
1	1	1.000 000	1.0000000	50	2 500	7.071 068	.02000000
2	4	1.414 214	.5000000	51	2 601	7.141 428	.01960784
3	9	1.732 051	.3333333	52	2 704	7.211 103	.01923077
4	16	2.000 000	.2500000	53	2 809	7.280 110	.01886792
5	25	2.236 068	.2000000	54	2 916	7.348 469	.01851852
6	36	2.449 490	.1666667	55	3 025	7.416 198	.01818182
7	49	2.645 751	.1428571	56	3 136	7.483 315	.01785714
8	64	2.828 427	.1250000	57	3 249	7.549 834	.01754386
9	81	3.000 000	.1111111	58	3 364	7.615 773	.01724138
10	100	3.162 278	.1000000	59	3 481	7.681 146	.01694915
11	121	3.316 625	.09090909	60	3 600	7.745 967	.01666667
12	144	3.464 102	.08333333	61	3 721	7.810 250	.01639344
13	169	3.605 551	.07692308	62	3 844	7.874 008	.01612903
14	196	3.741 657	.07142857	63	3 969	7.937 254	.01587302
15	225	3.872 983	.06666667	64	4 096	8.000 000	.01562500
16	256	4.000 000	.06250000	65	4 225	8.062 258	.01538462
17	289	4.123 106	.05882353	66	4 356	8.124 038	.01515152
18	324	4.242 641	.05555556	67	4 489	8.185 353	.01492537
19	361	4.358 899	.05263158	68	4 624	8.246 211	.01470588
20	400	4.472 136	.05000000	69	4 761	8.306 624	.01449275
21	441	4.582 576	.04761905	70	4 900	8.366 600	.01428571
22	484	4.690 416	.04545455	71	5 041	8.426 150	.01408451
23	529	4.795 832	.04347826	72	5 184	8.485 281	.01388889
24	576	4.898 979	.04166667	73	5 329	8.544 004	.01369863
25	625	5.000 000	.04000000	74	5 476	8.602 325	.01351351
26	676	5.099 020	.03846154	75	5 625	8.660 254	.01333333
27	729	5.196 152	.03703704	76	5 776	8.717 798	.01315789
28	784	5.291 503	.03571429	77	5 929	8.774 964	.01298701
29	841	5.385 165	.03448276	78	6 084	8.831 761	.01282051
30	900	5.477 226	.03333333	79	6 241	8.888 194	.01265823
31	961	5.567 764	.03225806	80	6 400	8.944 272	.01250000
32	1 024	5.656 854	.03125000	81	6 561	9.000 000	.01234568
33	1 089	5.744 563	.03030303	82	6 724	9.055 385	.01219512
34	1 156	5.830 952	.02941176	83	6 889	9.110 434	.01204819
35	1 225	5.916 080	.02857143	84	7 056	9.165 151	.01190476
36	1 296	6.000 000	.02777778	85	7 225	9.219 544	.01176471
37	1 369	6.082 763	.02702703	86	7 396	9.273 618	.01162791
38	1 444	6.164 414	.02631579	87	7 569	9.327 379	.01149425
39	1 521	6.244 998	.02564103	88	7 744	9.380 832	.01136364
40	1 600	6.324 555	.02500000	89	7 921	9.433 981	.01123596
41	1 681	6.403 124	.02439024	90	8 100	9.486 833	.01111111
42	1 764	6.480 741	.02380952	91	8 281	9.539 392	.01098901
43	1 849	6.557 439	.02325581	92	8 464	9.591 663	.01086957
44	1 936	6.633 250	.02272727	93	8 649	9.643 651	.01075269
45	2 025	6.708 204	.02222222	94	8 836	9.695 360	.01063830
46	2 116	6.782 330	.02173913	95	9 025	9.746 794	.01052632
47	2 209	6.855 655	.02127660	96	9 216	9.797 959	.01041667
48	2 304	6.928 203	.02083333	97	9 409	9.848 858	.01030928
49	2 401	7.000 000	.02040816	98	9 604	9.899 495	.01020408
50	2 500	7.071 068	.02000000	99	9 801	9.949 874	.01010101
				100	10 000	10.00000	.01000000



## Squares, Square Roots, and Reciprocals (cont.)

$n$	$n^2$	$\sqrt{n}$	$1/n$ .0	$n$	$n^2$	$\sqrt{n}$	$1/n$ .00
100	10 000	10.00000	10000000	150	22 500	12.24745	6666667
101	10 201	10.04988	09900990	151	22 801	12.28821	6622517
102	10 404	10.09950	09803922	152	23 104	12.32883	6578947
103	10 609	10.14889	09708738	153	23 409	12.36932	6535948
104	10 816	10.19804	09615385	154	23 716	12.40967	6493506
105	11 025	10.24695	09523810	155	24 025	12.44990	6451613
106	11 236	10.29563	09433962	156	24 336	12.49000	6410256
107	11 449	10.34408	09345794	157	24 649	12.52996	6369427
108	11 664	10.39230	09259259	158	24 964	12.56981	6329114
109	11 881	10.44031	09174312	159	25 281	12.60952	6289308
110	12 100	10.48809	09090909	160	25 600	12.64911	6250000
111	12 321	10.53565	09009009	161	25 921	12.68858	6211180
112	12 544	10.58301	08928571	162	26 244	12.72792	6172840
113	12 769	10.63015	08849558	163	26 569	12.76715	6134969
114	12 996	10.67708	08771930	164	26 896	12.80625	6097561
115	13 225	10.72381	08695652	165	27 225	12.84523	6060606
116	13 456	10.77033	08620690	166	27 556	12.88410	6024096
117	13 689	10.81665	08547009	167	27 889	12.92285	5988024
118	13 924	10.86278	08474576	168	28 224	12.96148	5952381
119	14 161	10.90871	08403361	169	28 561	13.00000	5917160
120	14 400	10.95445	08333333	170	28 900	13.03840	5882353
121	14 641	11.00000	08264463	171	29 241	13.07670	5847953
122	14 884	11.04536	08196721	172	29 584	13.11488	5813953
123	15 129	11.09064	08130081	173	29 929	13.15295	5780347
124	15 376	11.13553	08064516	174	30 276	13.19091	5747126
125	15 625	11.18034	08000000	175	30 625	13.22876	5714286
126	15 876	11.22497	07936508	176	30 976	13.26650	5681818
127	16 129	11.26943	07874016	177	31 329	13.30413	5649718
128	16 384	11.31371	07812500	178	31 684	13.34166	5617978
129	16 641	11.35782	07751938	179	32 041	13.37909	5586592
130	16 900	11.40175	07692308	180	32 400	13.41641	5555556
131	17 161	11.44552	07633588	181	32 761	13.45362	5524862
132	17 424	11.48913	07575758	182	33 124	13.49074	5494505
133	17 689	11.53256	07518797	183	33 489	13.52775	5464481
134	17 956	11.57584	07462687	184	33 856	13.56466	5434783
135	18 225	11.61895	07407407	185	34 225	13.60147	5405405
136	18 496	11.66190	07352941	186	34 596	13.63818	5376344
137	18 769	11.70470	07299270	187	34 969	13.67479	5347594
138	19 044	11.74734	07246377	188	35 344	13.71131	5319149
139	19 321	11.78983	07194245	189	35 721	13.74773	5291005
140	19 600	11.83216	07142857	190	36 100	13.78405	5263158
141	19 881	11.87434	07092199	191	36 481	13.82027	5235602
142	20 164	11.91638	07042254	192	36 864	13.85641	5208333
143	20 449	11.95826	06993007	193	37 249	13.89244	5181347
144	20 736	12.00000	06944444	194	37 636	13.92839	5154639
145	21 025	12.04159	06896552	195	38 025	13.96424	5128205
146	21 316	12.08305	06849315	196	38 416	14.00000	5102041
147	21 609	12.12436	06802721	197	38 809	14.03567	5076142
148	21 904	12.16553	06756757	198	39 204	14.07125	5050505
149	22 201	12.20656	06711409	199	39 601	14.10674	5025126
150	22 500	12.24745	06666667	200	40 000	14.14214	5000000

## Squares, Square Roots, and Reciprocals (cont.)

$n$	$n^2$	$\sqrt{n}$	$1/n$ .00	$n$	$n^2$	$\sqrt{n}$	$1/n$ .00
200	40 000	14 14214	5000000	250	62 500	15 81139	4000000
201	40 401	14.17745	4975124	251	63 001	15 84298	3984064
202	40 804	14 21267	4950495	252	63 504	15 87451	3968254
203	41 209	14 24781	4926108	253	64 009	15 90597	3952569
204	41 616	14 28286	4901961	254	64 516	15.93738	3937008
205	42 025	14 31732	4878049	255	65 025	15.96872	3921569
206	42 436	14.35270	4854369	256	65 536	16.00000	3906250
207	42 849	14 38749	4830918	257	66 049	16 03122	3891051
208	43 264	14.42221	4807692	258	66 564	16.06238	3875969
209	43 681	14.45683	4784689	259	67 081	16.09348	3861004
210	44 100	14.49138	4761905	260	67 600	16 12452	3846154
211	44 521	14 52584	4739336	261	68 121	16.15549	3831418
212	44 944	14.56022	4716981	262	68 644	16.18641	3816794
213	45 369	14.59452	4694836	263	69 169	16 21727	3802281
214	45 796	14.62874	4672897	264	69 696	16 24808	3787879
215	46 225	14 66288	4651163	265	70 225	16 27882	3773585
216	46 656	14 69694	4629630	266	70 756	16 30951	3759398
217	47 089	14 73092	4608295	267	71 289	16 34013	3745318
218	47 524	14.76482	4587156	268	71 824	16.37071	3731343
219	47 961	14 79865	4566210	269	72 361	16 40122	3717472
220	48 400	14 83240	4545455	270	72 900	16.43168	3703704
221	48 841	14.86607	4524887	271	73 441	16.46208	3690037
222	49 284	14.89966	4504505	272	73 984	16 49242	3676471
223	49 729	14.93318	4484305	273	74 529	16 52271	3663004
224	50 176	14.96663	4464286	274	75 076	16.55295	3649635
225	50 625	15.00000	4444444	275	75 625	16 58312	3636364
226	51 076	15 03330	4424779	276	76 176	16.61325	3623188
227	51 529	15.06652	4405286	277	76 729	16.64332	3610108
228	51 984	15.09967	4385965	278	77 284	16 67333	3597122
229	52 441	15.13275	4366812	279	77 841	16 70329	3584229
230	52 900	15.16575	4347826	280	78 400	16 73320	3571429
231	53 361	15.19868	4329004	281	78 961	16.76305	3558719
232	53 824	15.23155	4310345	282	79 524	16.79286	3546099
233	54 289	15.26434	4291845	283	80 089	16 82260	3533569
234	54 756	15.29706	4273504	284	80 656	16.85230	3521127
235	55 225	15 32971	4255319	285	81 225	16 88194	3508772
236	55 696	15.36229	4237288	286	81 796	16.91153	3496503
237	56 169	15.39480	4219409	287	82 369	16.94107	3484321
238	56 644	15.42725	4201681	288	82 944	16.97056	3472222
239	57 121	15.45962	4184100	289	83 521	17.00000	3460208
240	57 600	15.49193	4166667	290	84 100	17.02939	3448276
241	58 081	15 52417	4149378	291	84 681	17.05872	3436426
242	58 564	15.55635	4132231	292	85 264	17.08801	3424658
243	59 049	15.58846	4115226	293	85 849	17.11724	3412969
244	59 536	15.62050	4098361	294	86 436	17.14643	3401361
245	60 025	15.65248	4081633	295	87 025	17.17556	3389831
246	60 516	15.68439	4065041	296	87 616	17 20465	3378378
247	61 009	15.71623	4048583	297	88 209	17 23369	3367003
248	61 504	15.74802	4032258	298	88 804	17.26268	3355705
249	62 001	15.77973	4016064	299	89 401	17 29162	3344482
250	62 500	15.81139	4000000	300	90 000	17.32051	3333333

## Squares, Square Roots, and Reciprocals (cont.)

$n$	$n^2$	$\sqrt{n}$	$1/n$ .00	$n$	$n^2$	$\sqrt{n}$	$1/n$ .00
300	90 000	17.32051	3333333	350	122 500	18.70829	2857143
301	90 601	17.34935	3322259	351	123 201	18.73499	2849003
302	91 204	17.37815	3311258	352	123 904	18.76166	2840909
303	91 809	17.40690	3300380	353	124 609	18.78829	2832861
304	92 416	17.43560	3289474	354	125 316	18.81489	2824859
305	93 025	17.46425	3278689	355	126 025	18.84144	2816901
306	93 636	17.49286	3267974	356	126 736	18.86796	2808989
307	94 249	17.52142	3257329	357	127 449	18.89444	2801120
308	94 864	17.54993	3246753	358	128 164	18.92089	2793296
309	95 481	17.57840	3236246	359	128 881	18.94730	2785515
310	96 100	17.60682	3225806	360	129 600	18.97367	2777778
311	96 721	17.63519	3215434	361	130 321	19.00000	2770083
312	97 344	17.66352	3205128	362	131 044	19.02630	2762431
313	97 969	17.69181	3194888	363	131 769	19.05256	2754821
314	98 596	17.72005	3184713	364	132 496	19.07878	2747253
315	99 225	17.74824	3174603	365	133 225	19.10497	2739726
316	99 856	17.77639	3164557	366	133 956	19.13113	2732240
317	100 489	17.80449	3154574	367	134 689	19.15724	2724796
318	101 124	17.83255	3144654	368	135 424	19.18333	2717391
319	101 761	17.86057	3134796	369	136 161	19.20937	2710027
320	102 400	17.88854	3125000	370	136 900	19.23538	2702703
321	103 041	17.91647	3115265	371	137 641	19.26136	2695418
322	103 684	17.94436	3105590	372	138 384	19.28730	2688172
323	104 329	17.97220	3095975	373	139 129	19.31321	2680965
324	104 976	18.00000	3086420	374	139 876	19.33908	2673797
325	105 625	18.02776	3076923	375	140 625	19.36492	2666667
326	106 276	18.05547	3067485	376	141 376	19.39072	2659574
327	106 929	18.08314	3058104	377	142 129	19.41649	2652520
328	107 584	18.11077	3048780	378	142 884	19.44223	2645503
329	108 241	18.13836	3039514	379	143 641	19.46792	2638522
330	108 900	18.16590	3030303	380	144 400	19.49359	2631579
331	109 561	18.19341	3021148	381	145 161	19.51922	2624672
332	110 224	18.22087	3012048	382	145 924	19.54483	2617801
333	110 889	18.24829	3003003	383	146 689	19.57039	2610966
334	111 556	18.27567	2994012	384	147 456	19.59592	2604167
335	112 225	18.30301	2985075	385	148 225	19.62142	2597403
336	112 896	18.33030	2976190	386	148 996	19.64688	2590674
337	113 569	18.35756	2967359	387	149 769	19.67232	2583979
338	114 244	18.38478	2958580	388	150 544	19.69772	2577320
339	114 921	18.41195	2949853	389	151 321	19.72308	2570694
340	115 600	18.43909	2941176	390	152 100	19.74842	2564103
341	116 281	18.46619	2932551	391	152 881	19.77372	2557545
342	116 964	18.49324	2923977	392	153 664	19.79899	2551020
343	117 649	18.52026	2915452	393	154 449	19.82423	2544529
344	118 336	18.54724	2906977	394	155 236	19.84943	2538071
345	119 025	18.57418	2898551	395	156 025	19.87461	2531646
346	119 716	18.60108	2890173	396	156 816	19.89975	2525253
347	120 409	18.62794	2881844	397	157 609	19.92486	2518892
348	121 104	18.65476	2873563	398	158 404	19.94994	2512563
349	121 801	18.68154	2865330	399	159 201	19.97498	2506266
350	122 500	18.70829	2857143	400	160 000	20.00000	2500000

## Squares, Square Roots, and Reciprocals (cont.)

$n$	$n^2$	$\sqrt{n}$	$1/n$ .00	$n$	$n^2$	$\sqrt{n}$	$1/n$ .00
400	160 000	20.00000	2500000	450	202 500	21.21320	2222222
401	160 801	20.02498	2493766	451	203 401	21.23676	2217295
402	161 604	20.04994	2487562	452	204 304	21.26029	2212389
403	162 409	20.07486	2481390	453	205 209	21.28380	2207506
404	163 216	20.09975	2475248	454	206 116	21.30728	2202643
405	164 025	20.12461	2469136	455	207 025	21.33073	2197802
406	164 836	20.14944	2463054	456	207 936	21.35416	2192982
407	165 649	20.17424	2457002	457	208 849	21.37756	2188184
408	166 464	20.19901	2450980	458	209 764	21.40093	2183406
409	167 281	20.22375	2444988	459	210 681	21.42429	2178649
410	168 100	20.24846	2439024	460	211 600	21.44761	2173913
411	168 921	20.27313	2433090	461	212 521	21.47091	2169197
412	169 744	20.29778	2427184	462	213 444	21.49419	2164502
413	170 569	20.32240	2421308	463	214 369	21.51743	2159827
414	171 396	20.34699	2415459	464	215 296	21.54066	2155172
415	172 225	20.37155	2409639	465	216 225	21.56386	2150538
416	173 056	20.39608	2403846	466	217 156	21.58703	2145923
417	173 889	20.42058	2398082	467	218 089	21.61018	2141328
418	174 724	20.44505	2392344	468	219 024	21.63331	2136752
419	175 561	20.46949	2386635	469	219 961	21.65641	2132196
420	176 400	20.49390	2380952	470	220 900	21.67948	2127660
421	177 241	20.51828	2375297	471	221 841	21.70253	2123142
422	178 084	20.54264	2369668	472	222 784	21.72556	2118644
423	178 929	20.56696	2364066	473	223 729	21.74856	2114165
424	179 776	20.59126	2358491	474	224 676	21.77154	2109705
425	180 625	20.61553	2352941	475	225 625	21.79449	2105263
426	181 476	20.63977	2347418	476	226 576	21.81742	2100840
427	182 329	20.66398	2341920	477	227 529	21.84033	2096436
428	183 184	20.68816	2336449	478	228 484	21.86321	2092050
429	184 041	20.71232	2331002	479	229 441	21.88607	2087683
430	184 900	20.73644	2325581	480	230 400	21.90890	2083333
431	185 761	20.76054	2320186	481	231 361	21.93171	2079002
432	186 624	20.78461	2314815	482	232 324	21.95450	2074689
433	187 489	20.80865	2309469	483	233 289	21.97726	2070393
434	188 356	20.83267	2304147	484	234 256	22.00000	2066116
435	189 225	20.85665	2298851	485	235 225	22.02272	2061856
436	190 096	20.88061	2293578	486	236 196	22.04541	2057613
437	190 969	20.90454	2288330	487	237 169	22.06808	2053388
438	191 844	20.92845	2283105	488	238 144	22.09072	2049180
439	192 721	20.95233	2277904	489	239 121	22.11334	2044990
440	193 600	20.97618	2272727	490	240 100	22.13594	2040816
441	194 481	21.00000	2267574	491	241 081	22.15852	2036660
442	195 364	21.02380	2262443	492	242 064	22.18107	2032520
443	196 249	21.04757	2257336	493	243 049	22.20360	2028398
444	197 136	21.07131	2252252	494	244 036	22.22611	2024291
445	198 025	21.09502	2247191	495	245 025	22.24860	2020202
446	198 916	21.11871	2242152	496	246 016	22.27106	2016129
447	199 809	21.14237	2237136	497	247 009	22.29350	2012072
448	200 704	21.16601	2232143	498	248 004	22.31591	2008032
449	201 601	21.18962	2227171	499	249 001	22.33831	2004008
450	202 500	21.21320	2222222	500	250 000	22.36068	2000000

## Squares, Square Roots, and Reciprocals (cont.)

$n$	$n^2$	$\sqrt{n}$	$1/n$ .00	$n$	$n^2$	$\sqrt{n}$	$1/n$ .00
500	250 000	22.36068	2000000	550	302 500	23.45208	1818182
501	251 001	22.38303	1996008	551	303 601	23.47339	1814882
502	252 004	22.40536	1992032	552	304 704	23.49468	1811594
503	253 009	22.42766	1988072	553	305 809	23.51595	1808318
504	254 016	22.44994	1984127	554	306 916	23.53720	1805054
505	255 025	22.47221	1980198	555	308 025	23.55844	1801802
506	256 036	22.49444	1976285	556	309 136	23.57965	1798561
507	257 049	22.51666	1972387	557	310 249	23.60085	1795332
508	258 064	22.53886	1968504	558	311 364	23.62202	1792115
509	259 081	22.56103	1964637	559	312 481	23.64318	1788909
510	260 100	22.58318	1960784	560	313 600	23.66432	1785714
511	261 121	22.60531	1956947	561	314 721	23.68544	1782531
512	262 144	22.62742	1953125	562	315 844	23.70654	1779359
513	263 169	22.64950	1949318	563	316 969	23.72762	1776199
514	264 196	22.67157	1945525	564	318 096	23.74868	1773050
515	265 225	22.69361	1941748	565	319 225	23.76973	1769912
516	266 256	22.71563	1937984	566	320 356	23.79075	1766784
517	267 289	22.73763	1934236	567	321 489	23.81176	1763668
518	268 324	22.75961	1930502	568	322 624	23.83275	1760563
519	269 361	22.78157	1926782	569	323 761	23.85372	1757469
520	270 400	22.80351	1923077	570	324 900	23.87467	1754386
521	271 441	22.82542	1919386	571	326 041	23.89561	1751313
522	272 484	22.84732	1915709	572	327 184	23.91652	1748252
523	273 529	22.86919	1912046	573	328 329	23.93742	1745201
524	274 576	22.89105	1908397	574	329 476	23.95830	1742160
525	275 625	22.91288	1904762	575	330 625	23.97916	1739130
526	276 676	22.93469	1901141	576	331 776	24.00000	1736111
527	277 729	22.95648	1897533	577	332 929	24.02082	1733102
528	278 784	22.97825	1893939	578	334 084	24.04163	1730104
529	279 841	23.00000	1890359	579	335 241	24.06242	1727116
530	280 900	23.02173	1886792	580	336 400	24.08319	1724138
531	281 961	23.04344	1883239	581	337 561	24.10394	1721170
532	283 024	23.06513	1879699	582	338 724	24.12468	1718213
533	284 089	23.08679	1876173	583	339 889	24.14539	1715266
534	285 156	23.10844	1872659	584	341 056	24.16609	1712329
535	286 225	23.13007	1869159	585	342 225	24.18677	1709402
536	287 296	23.15167	1865672	586	343 396	24.20744	1706485
537	288 369	23.17326	1862197	587	344 569	24.22808	1703578
538	289 444	23.19483	1858736	588	345 744	24.24871	1700680
539	290 521	23.21637	1855288	589	346 921	24.26932	1697793
540	291 600	23.23790	1851852	590	348 100	24.28992	1694915
541	292 681	23.25941	1848429	591	349 281	24.31049	1692047
542	293 764	23.28089	1845018	592	350 464	24.33105	1689189
543	294 849	23.30236	1841621	593	351 649	24.35159	1686341
544	295 936	23.32381	1838236	594	352 836	24.37212	1683502
545	297 025	23.34524	1834862	595	354 025	24.39262	1680672
546	298 116	23.36664	1831502	596	355 216	24.41311	1677852
547	299 209	23.38803	1828154	597	356 409	24.43358	1675042
548	300 304	23.40940	1824818	598	357 604	24.45404	1672241
549	301 401	23.43075	1821494	599	358 801	24.47448	1669449
550	302 500	23.45208	1818182	600	360 000	24.49490	1666667

## Squares, Square Roots, and Reciprocals (cont.)

$n$	$n^2$	$\sqrt{n}$	$1/n$ .00	$n$	$n^2$	$\sqrt{n}$	$1/n$ .00
600	360 000	24.49490	1666667	650	422 500	25.49510	1538462
601	361 201	24.51530	1663894	651	423 801	25.51470	1536098
602	362 404	24.53569	1661130	652	425 104	25.53429	1533742
603	363 609	24.55606	1658375	653	426 409	25.55386	1531394
604	364 816	24.57641	1655629	654	427 716	25.57342	1529052
605	366 025	24.59675	1652893	655	429 025	25.59297	1526718
606	367 236	24.61707	1650165	656	430 336	25.61250	1524390
607	368 449	24.63737	1647446	657	431 649	25.63201	1522070
608	369 664	24.65766	1644737	658	432 964	25.65151	1519757
609	370 881	24.67793	1642036	659	434 281	25.67100	1517451
610	372 100	24.69818	1639344	660	435 600	25.69047	1515152
611	373 321	24.71841	1636661	661	436 921	25.70992	1512859
612	374 544	24.73863	1633987	662	438 244	25.72936	1510574
613	375 769	24.75884	1631321	663	439 569	25.74879	1508296
614	376 996	24.77902	1628664	664	440 896	25.76820	1506024
615	378 225	24.79919	1626016	665	442 225	25.78759	1503759
616	379 456	24.81935	1623377	666	443 556	25.80698	1501502
617	380 689	24.83948	1620746	667	444 889	25.82634	1499250
618	381 924	24.85961	1618123	668	446 224	25.84570	1497006
619	383 161	24.87971	1615509	669	447 561	25.86503	1494768
620	384 400	24.89980	1612903	670	448 900	25.88436	1492537
621	385 641	24.91987	1610306	671	450 241	25.90367	1490313
622	386 884	24.93993	1607717	672	451 584	25.92296	1488095
623	388 129	24.95997	1605136	673	452 929	25.94224	1485884
624	389 376	24.97999	1602564	674	454 276	25.96151	1483680
625	390 625	25.00000	1600000	675	455 625	25.98076	1481481
626	391 876	25.01999	1597444	676	456 976	26.00000	1479290
627	393 129	25.03997	1594896	677	458 329	26.01922	1477105
628	394 384	25.05993	1592357	678	459 684	26.03843	1474926
629	395 641	25.07987	1589825	679	461 041	26.05763	1472754
630	396 900	25.09980	1587302	680	462 400	26.07681	1470588
631	398 161	25.11971	1584786	681	463 761	26.09598	1468429
632	399 424	25.13961	1582278	682	465 124	26.11513	1466276
633	400 689	25.15949	1579779	683	466 489	26.13427	1464129
634	401 956	25.17936	1577287	684	467 856	26.15339	1461988
635	403 225	25.19921	1574803	685	469 225	26.17250	1459854
636	404 496	25.21904	1572327	686	470 596	26.19160	1457726
637	405 769	25.23886	1569859	687	471 969	26.21068	1455604
638	407 044	25.25866	1567398	688	473 344	26.22975	1453488
639	408 321	25.27845	1564945	689	474 721	26.24881	1451379
640	409 600	25.29822	1562500	690	476 100	26.26785	1449275
641	410 881	25.31798	1560062	691	477 481	26.28688	1447178
642	412 164	25.33772	1557632	692	478 864	26.30589	1445087
643	413 449	25.35744	1555210	693	480 249	26.32489	1443001
644	414 736	25.37716	1552795	694	481 636	26.34388	1440922
645	416 025	25.39685	1550388	695	483 025	26.36285	1438849
646	417 316	25.41653	1547988	696	484 416	26.38181	1436782
647	418 609	25.43619	1545595	697	485 809	26.40076	1434720
648	419 904	25.45584	1543210	698	487 204	26.41969	1432665
649	421 201	25.47548	1540832	699	488 601	26.43861	1430615
650	422 500	25.49510	1538462	700	490 000	26.45751	1428571

## Squares, Square Roots, and Reciprocals (cont.)

$n$	$n^2$	$\sqrt{n}$	$1/n$ .00	$n$	$n^2$	$\sqrt{n}$	$1/n$ .00
700	490 000	26.45751	1428571	750	562 500	27.38613	1333333
701	491 401	26.47640	1426534	751	564 001	27.40438	1331558
702	492 804	26.49528	1424501	752	565 504	27.42262	1329787
703	494 209	26.51415	1422475	753	567 009	27.44085	1328021
704	495 616	26.53300	1420455	754	568 516	27.45906	1326260
705	497 025	26.55184	1418440	755	570 025	27.47726	1324503
706	498 436	26.57066	1416431	756	571 536	27.49545	1322751
707	499 849	26.58947	1414427	757	573 049	27.51363	1321004
708	501 264	26.60827	1412429	758	574 564	27.53180	1319261
709	502 681	26.62705	1410437	759	576 081	27.54995	1317523
710	504 100	26.64583	1408451	760	577 600	27.56810	1315789
711	505 521	26.66458	1406470	761	579 121	27.58623	1314060
712	506 944	26.68333	1404494	762	580 644	27.60435	1312336
713	508 369	26.70206	1402525	763	582 169	27.62245	1310616
714	509 796	26.72078	1400560	764	583 696	27.64055	1308901
715	511 225	26.73948	1398601	765	585 225	27.65863	1307190
716	512 656	26.75818	1396648	766	586 756	27.67671	1305483
717	514 089	26.77686	1394700	767	588 289	27.69476	1303781
718	515 524	26.79552	1392758	768	589 824	27.71281	1302083
719	516 961	26.81418	1390821	769	591 361	27.73085	1300390
720	518 400	26.83282	1388889	770	592 900	27.74887	1298701
721	519 841	26.85144	1386963	771	594 441	27.76689	1297017
722	521 284	26.87006	1385042	772	595 984	27.78490	1295337
723	522 729	26.88866	1383126	773	597 529	27.80288	1293661
724	524 176	26.90725	1381215	774	599 076	27.82086	1291990
725	525 625	26.92582	1379310	775	600 625	27.83882	1290323
726	527 076	26.94439	1377410	776	602 176	27.85678	1288660
727	528 529	26.96294	1375516	777	603 729	27.87472	1287001
728	529 984	26.98148	1373626	778	605 284	27.89265	1285347
729	531 441	27.00000	1371742	779	606 841	27.91057	1283697
730	532 900	27.01851	1369863	780	608 400	27.92848	1282051
731	534 361	27.03701	1367989	781	609 961	27.94638	1280410
732	535 824	27.05550	1366120	782	611 524	27.96426	1278772
733	537 289	27.07397	1364256	783	613 089	27.98214	1277139
734	538 756	27.09243	1362398	784	614 656	28.00000	1275510
735	540 225	27.11088	1360544	785	616 225	28.01785	1273885
736	541 696	27.12932	1358696	786	617 796	28.03569	1272265
737	543 169	27.14774	1356852	787	619 369	28.05352	1270648
738	544 644	27.16616	1355014	788	620 944	28.07134	1269036
739	546 121	27.18455	1353180	789	622 521	28.08914	1267427
740	547 600	27.20294	1351351	790	624 100	28.10694	1265823
741	549 081	27.22132	1349528	791	625 681	28.12472	1264223
742	550 564	27.23968	1347709	792	627 264	28.14249	1262626
743	552 049	27.25803	1345895	793	628 849	28.16026	1261034
744	553 536	27.27636	1344086	794	630 436	28.17801	1259446
745	555 025	27.29469	1342282	795	632 025	28.19574	1257862
746	556 516	27.31300	1340483	796	633 616	28.21347	1256281
747	558 009	27.33130	1338688	797	635 209	28.23119	1254705
748	559 504	27.34959	1336898	798	636 804	28.24889	1253133
749	561 001	27.36786	1335113	799	638 401	28.26659	1251564
750	562 500	27.38613	1333333	800	640 000	28.28427	1250000

## Squares, Square Roots, and Reciprocals (cont.)

$n$	$n^2$	$\sqrt{n}$	$1/n$ .00	$n$	$n^2$	$\sqrt{n}$	$1/n$ .00
800	640 000	28.28427	1250000	850	722 500	29.15476	1176471
801	641 601	28.30194	1248439	851	724 201	29.17190	1175088
802	643 204	28.31960	1246883	852	725 904	29.18904	1173709
803	644 809	28.33725	1245330	853	727 609	29.20616	1172333
804	646 416	28.35489	1243781	854	729 316	29.22328	1170960
805	648 025	28.37252	1242236	855	731 025	29.24038	1169591
806	649 636	28.39014	1240695	856	732 736	29.25748	1168224
807	651 249	28.40775	1239157	857	734 449	29.27456	1166861
808	652 864	28.42534	1237624	858	736 164	29.29164	1165501
809	654 481	28.44293	1236094	859	737 881	29.30870	1164144
810	656 100	28.46050	1234568	860	739 600	29.32576	1162791
811	657 721	28.47806	1233046	861	741 321	29.34280	1161440
812	659 344	28.49561	1231527	862	743 044	29.35984	1160093
813	660 969	28.51315	1230012	863	744 769	29.37686	1158749
814	662 596	28.53069	1228501	864	746 496	29.39388	1157407
815	664 225	28.54820	1226994	865	748 225	29.41088	1156069
816	665 856	28.56571	1225490	866	749 956	29.42788	1154734
817	667 489	28.58321	1223990	867	751 689	29.44486	1153403
818	669 124	28.60070	1222494	868	753 424	29.46184	1152074
819	670 761	28.61818	1221001	869	755 161	29.47881	1150748
820	672 400	28.63564	1219512	870	756 900	29.49576	1149425
821	674 041	28.65310	1218027	871	758 641	29.51271	1148106
822	675 684	28.67054	1216545	872	760 384	29.52965	1146789
823	677 329	28.68798	1215067	873	762 129	29.54657	1145475
824	678 976	28.70540	1213592	874	763 876	29.56349	1144165
825	680 625	28.72281	1212121	875	765 625	29.58040	1142857
826	682 276	28.74022	1210654	876	767 376	29.59730	1141553
827	683 929	28.75761	1209190	877	769 129	29.61419	1140251
828	685 584	28.77499	1207729	878	770 884	29.63106	1138952
829	687 241	28.79236	1206273	879	772 641	29.64793	1137656
830	688 900	28.80972	1204819	880	774 400	29.66479	1136364
831	690 561	28.82707	1203369	881	776 161	29.68164	1135074
832	692 224	28.84441	1201923	882	777 924	29.69848	1133787
833	693 889	28.86174	1200480	883	779 689	29.71532	1132503
834	695 556	28.87906	1199041	884	781 456	29.73214	1131222
835	697 225	28.89637	1197605	885	783 225	29.74895	1129944
836	698 896	28.91366	1196172	886	784 996	29.76575	1128668
837	700 569	28.93095	1194743	887	786 769	29.78255	1127396
838	702 244	28.94823	1193317	888	788 544	29.79933	1126126
839	703 921	28.96550	1191895	889	790 321	29.81610	1124859
840	705 600	28.98275	1190476	890	792 100	29.83287	1123596
841	707 281	29.00000	1189061	891	793 881	29.84962	1122334
842	708 964	29.01724	1187648	892	795 664	29.86637	1121076
843	710 649	29.03446	1186240	893	797 449	29.88311	1119821
844	712 336	29.05168	1184834	894	799 236	29.89983	1118568
845	714 025	29.06888	1183432	895	801 025	29.91655	1117318
846	715 716	29.08608	1182033	896	802 816	29.93326	1116071
847	717 409	29.10326	1180638	897	804 609	29.94996	1114827
848	719 104	29.12044	1179245	898	806 404	29.96665	1113586
849	720 801	29.13760	1177856	899	808 201	29.98333	1112347
850	722 500	29.15476	1176471	900	810 000	30.00000	1111111



**Squares, Square Roots, and Reciprocals (cont.)**

$n$	$n^2$	$\sqrt{n}$	$1/n$ .00	$n$	$n^2$	$\sqrt{n}$	$1/n$ .00
900	810 000	30.00000	1111111	950	902 500	30.82207	1052632
901	811 801	30.01666	1109878	951	904 401	30.83829	1051525
902	813 604	30.03331	1108647	952	906 304	30.85450	1050420
903	815 409	30.04996	1107420	953	908 209	30.87070	1049318
904	817 216	30.06659	1106195	954	910 116	30.88689	1048218
905	819 025	30.08322	1104972	955	912 025	30.90307	1047120
906	820 836	30.09983	1103753	956	913 936	30.91925	1046025
907	822 649	30.11644	1102536	957	915 849	30.93542	1044932
908	824 464	30.13304	1101322	958	917 764	30.95158	1043841
909	826 281	30.14963	1100110	959	919 681	30.96773	1042753
910	828 100	30.16621	1098901	960	921 600	30.98387	1041667
911	829 921	30.18278	1097695	961	923 521	31.00000	1040583
912	831 744	30.19934	1096491	962	925 444	31.01612	1039501
913	833 569	30.21589	1095290	963	927 369	31.03224	1038422
914	835 396	30.23243	1094092	964	929 296	31.04835	1037344
915	837 225	30.24897	1092896	965	931 225	31.06445	1036269
916	839 056	30.26549	1091703	966	933 156	31.08054	1035197
917	840 889	30.28201	1090513	967	935 089	31.09662	1034126
918	842 724	30.29851	1089325	968	937 024	31.11270	1033058
919	844 561	30.31501	1088139	969	938 961	31.12876	1031992
920	846 400	30.33150	1086957	970	940 900	31.14482	1030928
921	848 241	30.34798	1085776	971	942 841	31.16087	1029866
922	850 084	30.36445	1084599	972	944 784	31.17691	1028807
923	851 929	30.38092	1083424	973	946 729	31.19295	1027749
924	853 776	30.39737	1082251	974	948 676	31.20897	1026694
925	855 625	30.41381	1081081	975	950 625	31.22499	1025641
926	857 476	30.43025	1079914	976	952 576	31.24100	1024590
927	859 329	30.44667	1078749	977	954 529	31.25700	1023541
928	861 184	30.46309	1077586	978	956 484	31.27299	1022495
929	863 041	30.47950	1076426	979	958 441	31.28898	1021450
930	864 900	30.49590	1075269	980	960 400	31.30495	1020408
931	866 761	30.51229	1074114	981	962 361	31.32092	1019368
932	868 624	30.52868	1072961	982	964 324	31.33688	1018330
933	870 489	30.54505	1071811	983	966 289	31.35283	1017294
934	872 356	30.56141	1070664	984	968 256	31.36877	1016260
935	874 225	30.57777	1069519	985	970 225	31.38471	1015228
936	876 096	30.59412	1068376	986	972 196	31.40064	1014199
937	877 969	30.61046	1067236	987	974 169	31.41656	1013171
938	879 844	30.62679	1066098	988	976 144	31.43247	1012146
939	881 721	30.64311	1064963	989	978 121	31.44837	1011122
940	883 600	30.65942	1063830	990	980 100	31.46427	1010101
941	885 481	30.67572	1062699	991	982 081	31.48015	1009082
942	887 364	30.69202	1061571	992	984 064	31.49603	1008065
943	889 249	30.70831	1060445	993	986 049	31.51190	1007049
944	891 136	30.72458	1059322	994	988 036	31.52777	1006036
945	893 025	30.74085	1058201	995	990 025	31.54362	1005025
946	894 916	30.75711	1057082	996	992 016	31.55947	1004016
947	896 809	30.77337	1055966	997	994 009	31.57531	1003009
948	898 704	30.78961	1054852	998	996 004	31.59114	1002004
949	900 601	30.80584	1053741	999	998 001	31.60696	1001001
950	902 500	30.82207	1052632	1000	1 000 000	31.62278	1000000

**Squares, Square Roots, and Reciprocals (cont.)**

$n$	$\sqrt{n}$	$n$	$\sqrt{n}$	$n$	$\sqrt{n}$	$n$	$\sqrt{n}$
1000	31.62278	1500	38.72983	2000	44.72136	2500	50.00000
1010	31.78050	1510	38.85872	2010	44.83302	2510	50.09990
1020	31.93744	1520	38.98718	2020	44.94441	2520	50.19960
1030	32.09361	1530	39.11521	2030	45.05552	2530	50.29911
1040	32.24903	1540	39.24283	2040	45.16636	2540	50.39841
1050	32.40370	1550	39.37004	2050	45.27693	2550	50.49752
1060	32.55764	1560	39.49684	2060	45.38722	2560	50.59644
1070	32.71085	1570	39.62323	2070	45.49725	2570	50.69517
1080	32.86335	1580	39.74921	2080	45.60702	2580	50.79370
1090	33.01515	1590	39.87480	2090	45.71652	2590	50.89204
1100	33.16625	1600	40.00000	2100	45.82576	2600	50.99020
1110	33.31666	1610	40.12481	2110	45.93474	2610	51.08816
1120	33.46640	1620	40.24922	2120	46.04346	2620	51.18594
1130	33.61547	1630	40.37326	2130	46.15192	2630	51.28353
1140	33.76389	1640	40.49691	2140	46.26013	2640	51.38093
1150	33.91165	1650	40.62019	2150	46.36809	2650	51.47815
1160	34.05877	1660	40.74310	2160	46.47580	2660	51.57519
1170	34.20526	1670	40.86563	2170	46.58326	2670	51.67204
1180	34.35113	1680	40.98780	2180	46.69047	2680	51.76872
1190	34.49638	1690	41.10961	2190	46.79744	2690	51.86521
1200	34.64102	1700	41.23106	2200	46.90416	2700	51.96152
1210	34.78505	1710	41.35215	2210	47.01064	2710	52.05766
1220	34.92850	1720	41.47288	2220	47.11688	2720	52.15362
1230	35.07136	1730	41.59327	2230	47.22288	2730	52.24940
1240	35.21363	1740	41.71331	2240	47.32864	2740	52.34501
1250	35.35534	1750	41.83300	2250	47.43416	2750	52.44044
1260	35.49648	1760	41.95235	2260	47.53946	2760	52.53570
1270	35.63706	1770	42.07137	2270	47.64452	2770	52.63079
1280	35.77709	1780	42.19005	2280	47.74935	2780	52.72571
1290	35.91657	1790	42.30839	2290	47.85394	2790	52.82045
1300	36.05551	1800	42.42641	2300	47.95832	2800	52.91503
1310	36.19392	1810	42.54409	2310	48.06246	2810	53.00943
1320	36.33180	1820	42.66146	2320	48.16638	2820	53.10367
1330	36.46917	1830	42.77850	2330	48.27007	2830	53.19774
1340	36.60601	1840	42.89522	2340	48.37355	2840	53.29165
1350	36.74235	1850	43.01163	2350	48.47680	2850	53.38539
1360	36.87818	1860	43.12772	2360	48.57983	2860	53.47897
1370	37.01351	1870	43.24350	2370	48.68265	2870	53.57238
1380	37.14835	1880	43.35897	2380	48.78524	2880	53.66563
1390	37.28270	1890	43.47413	2390	48.88763	2890	53.75872
1400	37.41657	1900	43.58899	2400	48.98979	2900	53.85165
1410	37.54997	1910	43.70355	2410	49.09175	2910	53.94442
1420	37.68289	1920	43.81780	2420	49.19350	2920	54.03702
1430	37.81534	1930	43.93177	2430	49.29503	2930	54.12947
1440	37.94733	1940	44.04543	2440	49.39636	2940	54.22177
1450	38.07887	1950	44.15880	2450	49.49747	2950	54.31390
1460	38.20995	1960	44.27189	2460	49.59839	2960	54.40588
1470	38.34058	1970	44.38468	2470	49.69909	2970	54.49771
1480	38.47077	1980	44.49719	2480	49.79960	2980	54.58938
1490	38.60052	1990	44.60942	2490	49.89990	2990	54.68089
1500	38.72983	2000	44.72136	2500	50.00000	3000	54.77226

## Squares, Square Roots, and Reciprocals (cont.)

$n$	$\sqrt{n}$	$n$	$\sqrt{n}$	$n$	$\sqrt{n}$	$n$	$\sqrt{n}$
3000	54.77226	3500	59.16080	4000	63.24555	4500	67.08204
3010	54.8347	3510	59.24525	4010	63.32456	4510	67.15653
3020	54.95463	3520	59.32959	4020	63.40347	4520	67.23095
3030	55.04544	3530	59.41380	4030	63.48228	4530	67.30527
3040	55.13620	3540	59.49790	4040	63.56099	4540	67.37952
3050	55.22681	3550	59.58188	4050	63.63961	4550	67.45369
3060	55.31727	3560	59.66574	4060	63.71813	4560	67.52777
3070	55.40758	3570	59.74948	4070	63.79655	4570	67.60178
3080	55.49775	3580	59.83310	4080	63.87488	4580	67.67570
3090	55.58777	3590	59.91661	4090	63.95311	4590	67.74954
3100	55.67764	3600	60.00000	4100	64.03124	4600	67.82330
3110	55.76737	3610	60.08328	4110	64.10928	4610	67.89698
3120	55.85696	3620	60.16644	4120	64.18723	4620	67.97058
3130	55.94640	3630	60.24948	4130	64.26508	4630	68.04410
3140	56.03570	3640	60.33241	4140	64.34283	4640	68.11755
3150	56.12486	3650	60.41523	4150	64.42049	4650	68.19091
3160	56.21388	3660	60.49793	4160	64.49806	4660	68.26419
3170	56.30275	3670	60.58052	4170	64.57554	4670	68.33740
3180	56.39149	3680	60.66300	4180	64.65292	4680	68.41053
3190	56.48008	3690	60.74537	4190	64.73021	4690	68.48357
3200	56.56854	3700	60.82763	4200	64.80741	4700	68.55655
3210	56.65686	3710	60.90977	4210	64.88451	4710	68.62944
3220	56.74504	3720	60.99180	4220	64.96153	4720	68.70226
3230	56.83309	3730	61.07373	4230	65.03845	4730	68.77500
3240	56.92100	3740	61.15554	4240	65.11528	4740	68.84766
3250	57.00877	3750	61.23724	4250	65.19202	4750	68.92024
3260	57.09641	3760	61.31884	4260	65.26868	4760	68.99275
3270	57.18391	3770	61.40033	4270	65.34524	4770	69.06519
3280	57.27128	3780	61.48170	4280	65.42171	4780	69.13754
3290	57.35852	3790	61.56298	4290	65.49809	4790	69.20983
3300	57.44563	3800	61.64414	4300	65.57439	4800	69.28203
3310	57.53260	3810	61.72520	4310	65.65059	4810	69.35416
3320	57.61944	3820	61.80615	4320	65.72671	4820	69.42622
3330	57.70615	3830	61.88699	4330	65.80274	4830	69.49820
3340	57.79273	3840	61.96778	4340	65.87868	4840	69.57011
3350	57.87918	3850	62.04837	4350	65.95453	4850	69.64194
3360	57.96551	3860	62.12890	4360	66.03030	4860	69.71370
3370	58.05170	3870	62.20932	4370	66.10598	4870	69.78539
3380	58.13777	3880	62.28965	4380	66.18157	4880	69.85700
3390	58.22371	3890	62.36986	4390	66.25708	4890	69.92853
3400	58.30952	3900	62.44998	4400	66.33250	4900	70.00000
3410	58.39521	3910	62.52999	4410	66.40783	4910	70.07139
3420	58.48077	3920	62.60990	4420	66.48308	4920	70.14271
3430	58.56620	3930	62.68971	4430	66.55825	4930	70.21396
3440	58.65151	3940	62.76942	4440	66.63332	4940	70.28513
3450	58.73670	3950	62.84903	4450	66.70832	4950	70.35624
3460	58.82176	3960	62.92853	4460	66.78323	4960	70.42727
3470	58.90671	3970	63.00794	4470	66.85806	4970	70.49823
3480	58.99152	3980	63.08724	4480	66.93280	4980	70.56912
3490	59.07622	3990	63.16645	4490	67.00746	4990	70.63993
3500	59.16080	4000	63.24555	4500	67.08204	5000	70.71068

**Squares, Square Roots, and Reciprocals (cont.)**

$n$	$\sqrt{n}$	$n$	$\sqrt{n}$	$n$	$\sqrt{n}$	$n$	$\sqrt{n}$
5000	70.71068	5500	74.16198	6000	77.45967	6500	80.62258
5010	70.78135	5510	74.22937	6010	77.52419	6510	80.68457
5020	70.85196	5520	74.29670	6020	77.58866	6520	80.74652
5030	70.92249	5530	74.36397	6030	77.65307	6530	80.80842
5040	70.99296	5540	74.43118	6040	77.71744	6540	80.87027
5050	71.06335	5550	74.49832	6050	77.78175	6550	80.93207
5060	71.13368	5560	74.56541	6060	77.84600	6560	80.99383
5070	71.20393	5570	74.63243	6070	77.91020	6570	81.05554
5080	71.27412	5580	74.69940	6080	77.97435	6580	81.11720
5090	71.34424	5590	74.76630	6090	78.03846	6590	81.17881
5100	71.41428	5600	74.83315	6100	78.10250	6600	81.24038
5110	71.48426	5610	74.89993	6110	78.16649	6610	81.30191
5120	71.55418	5620	74.96666	6120	78.23043	6620	81.36338
5130	71.62402	5630	75.03333	6130	78.29432	6630	81.42481
5140	71.69379	5640	75.09993	6140	78.35815	6640	81.48620
5150	71.76350	5650	75.16648	6150	78.42194	6650	81.54753
5160	71.83314	5660	75.23297	6160	78.48567	6660	81.60882
5170	71.90271	5670	75.29940	6170	78.54935	6670	81.67007
5180	71.97222	5680	75.36577	6180	78.61298	6680	81.73127
5190	72.04165	5690	75.43209	6190	78.67655	6690	81.79242
5200	72.11103	5700	75.49834	6200	78.74008	6700	81.85353
5210	72.18033	5710	75.56454	6210	78.80355	6710	81.91459
5220	72.24957	5720	75.63068	6220	78.86698	6720	81.97561
5230	72.31874	5730	75.69676	6230	78.93035	6730	82.03658
5240	72.38784	5740	75.76279	6240	78.99367	6740	82.09750
5250	72.45688	5750	75.82875	6250	79.05694	6750	82.15838
5260	72.52586	5760	75.89466	6260	79.12016	6760	82.21922
5270	72.59477	5770	75.96052	6270	79.18333	6770	82.28001
5280	72.66361	5780	76.02631	6280	79.24645	6780	82.34076
5290	72.73239	5790	76.09205	6290	79.30952	6790	82.40146
5300	72.80110	5800	76.15773	6300	79.37254	6800	82.46211
5310	72.86975	5810	76.22336	6310	79.43551	6810	82.42272
5320	72.93833	5820	76.28892	6320	79.49843	6820	82.48329
5330	73.00685	5830	76.35444	6330	79.56130	6830	82.54381
5340	73.07530	5840	76.41989	6340	79.62412	6840	82.70429
5350	73.14369	5850	76.48529	6350	79.68689	6850	82.76473
5360	73.21202	5860	76.55064	6360	79.74961	6860	82.82512
5370	73.28028	5870	76.61593	6370	79.81228	6870	82.88546
5380	73.34848	5880	76.68116	6380	79.87490	6880	82.94577
5390	73.41662	5890	76.74634	6390	79.93748	6890	83.00602
5400	73.48469	5900	76.81146	6400	80.00000	6900	83.06624
5410	73.55270	5910	76.87652	6410	80.06248	6910	83.12641
5420	73.62065	5920	76.94154	6420	80.12490	6920	83.18654
5430	73.68853	5930	77.00649	6430	80.18728	6930	83.24662
5440	73.75636	5940	77.07140	6440	80.24961	6940	83.30666
5450	73.82412	5950	77.13624	6450	80.31189	6950	83.36666
5460	73.89181	5960	77.20104	6460	80.37413	6960	83.42661
5470	73.95945	5970	77.26578	6470	80.43631	6970	83.48653
5480	74.02702	5980	77.33046	6480	80.49845	6980	83.54639
5490	74.09453	5990	77.39509	6490	80.56054	6990	83.60622
5500	74.16198	6000	77.45967	6500	80.62258	7000	83.66600

## Squares, Square Roots, and Reciprocals (cont.)

$n$	$\sqrt{n}$	$n$	$\sqrt{n}$	$n$	$\sqrt{n}$	$n$	$\sqrt{n}$
7000	83.66600	7500	86.60254	8000	89.44272	8500	92.19544
7010	83.72574	7510	86.66026	8010	89.49860	8510	92.24966
7020	83.78544	7520	86.71793	8020	89.55445	8520	92.30385
7030	83.84510	7530	86.77557	8030	89.61027	8530	92.35800
7040	83.90471	7540	86.83317	8040	89.66605	8540	92.41212
7050	83.96428	7550	86.89074	8050	89.72179	8550	92.46621
7060	84.02381	7560	86.94826	8060	89.77750	8560	92.52027
7070	84.08329	7570	87.00575	8070	89.83318	8570	92.57429
7080	84.14274	7580	87.06320	8080	89.88882	8580	92.62829
7090	84.20214	7590	87.12061	8090	89.94443	8590	92.68225
7100	84.26150	7600	87.17798	8100	90.00000	8600	92.73618
7110	84.32082	7610	87.23531	8110	90.05554	8610	92.79009
7120	84.38009	7620	87.29261	8120	90.11104	8620	92.84396
7130	84.43933	7630	87.34987	8130	90.16651	8630	92.89779
7140	84.49852	7640	87.40709	8140	90.22195	8640	92.95160
7150	84.55767	7650	87.46428	8150	90.27735	8650	93.00538
7160	84.61678	7660	87.52143	8160	90.33272	8660	93.05912
7170	84.67585	7670	87.57854	8170	90.38805	8670	93.11283
7180	84.73488	7680	87.63561	8180	90.44335	8680	93.16652
7190	84.79387	7690	87.69265	8190	90.49862	8690	93.22017
7200	84.85281	7700	87.74964	8200	90.55385	8700	93.27379
7210	84.91172	7710	87.80661	8210	90.60905	8710	93.32738
7220	84.97058	7720	87.86353	8220	90.66422	8720	93.38094
7230	85.02941	7730	87.92042	8230	90.71935	8730	93.43447
7240	85.08819	7740	87.97727	8240	90.77445	8740	93.48797
7250	85.14693	7750	88.03408	8250	90.82951	8750	93.54143
7260	85.20563	7760	88.09086	8260	90.88454	8760	93.59487
7270	85.26429	7770	88.14760	8270	90.93954	8770	93.64828
7280	85.32292	7780	88.20431	8280	90.99451	8780	93.70165
7290	85.38150	7790	88.26098	8290	91.04944	8790	93.75500
7300	85.44004	7800	88.31761	8300	91.10434	8800	93.80832
7310	85.49854	7810	88.37420	8310	91.15920	8810	93.86160
7320	85.55700	7820	88.43076	8320	91.21403	8820	93.91486
7330	85.61542	7830	88.48729	8330	91.26883	8830	93.96808
7340	85.67380	7840	88.54377	8340	91.32360	8840	94.02127
7350	85.73214	7850	88.60023	8350	91.37833	8850	94.07444
7360	85.79044	7860	88.65664	8360	91.43304	8860	94.12757
7370	85.84870	7870	88.71302	8370	91.48770	8870	94.18068
7380	85.90693	7880	88.76936	8380	91.54234	8880	94.23375
7390	85.96511	7890	88.82567	8390	91.59694	8890	94.28680
7400	86.02325	7900	88.88194	8400	91.65151	8900	94.33981
7410	86.08136	7910	88.93818	8410	91.70605	8910	94.39280
7420	86.13942	7920	88.99438	8420	91.76056	8920	94.44575
7430	86.19745	7930	89.05055	8430	91.81503	8930	94.49868
7440	86.25543	7940	89.10668	8440	91.86947	8940	94.55157
7450	86.31338	7950	89.16277	8450	91.92388	8950	94.60444
7460	86.37129	7960	89.21883	8460	91.97826	8960	94.65728
7470	86.42916	7970	89.27486	8470	92.03260	8970	94.71008
7480	86.48699	7980	89.33085	8480	92.08692	8980	94.76286
7490	86.54479	7990	89.38680	8490	92.14120	8990	94.81561
7500	86.60254	8000	89.44272	8500	92.19544	9000	94.86833

**Squares, Square Roots, and Reciprocals (cont.)**

$n$	$\sqrt{n}$	$n$	$\sqrt{n}$
9000	94.86833	9500	97.46794
9010	94.92102	9510	97.51923
9020	94.97368	9520	97.57049
9030	95.02631	9530	97.62172
9040	95.07891	9540	97.67292
9050	95.13149	9550	97.72410
9060	95.18403	9560	97.77525
9070	95.23655	9570	97.82638
9080	95.28903	9580	97.87747
9090	95.34149	9590	97.92855
9100	95.39392	9600	97.97959
9110	95.44632	9610	98.03061
9120	95.49869	9620	98.08160
9130	95.55103	9630	98.13256
9140	95.60335	9640	98.18350
9150	95.65563	9650	98.23441
9160	95.70789	9660	98.28530
9170	95.76012	9670	98.33616
9180	95.81232	9680	98.38699
9190	95.86449	9690	98.43780
9200	95.91663	9700	98.48858
9210	95.96874	9710	98.53933
9220	96.02083	9720	98.59006
9230	96.07289	9730	98.64076
9240	96.12492	9740	98.69144
9250	96.17692	9750	98.74209
9260	96.22889	9760	98.79271
9270	96.28084	9770	98.84331
9280	96.33276	9780	98.89388
9290	96.38465	9790	98.94443
9300	96.43561	9800	98.99495
9310	96.48634	9810	99.04544
9320	96.54015	9820	99.09591
9330	96.59193	9830	99.14636
9340	96.64368	9840	99.19677
9350	96.69540	9850	99.24717
9360	96.74709	9860	99.29753
9370	96.79876	9870	99.34787
9380	96.85040	9880	99.39819
9390	96.90201	9890	99.44848
9400	96.95360	9900	99.49874
9410	97.00515	9910	99.54898
9420	97.05668	9920	99.59920
9430	97.10819	9930	99.64939
9440	97.15966	9940	99.69955
9450	97.21111	9950	99.74969
9460	97.26253	9960	99.79980
9470	97.31393	9970	99.84989
9480	97.36529	9980	99.89995
9490	97.41663	9990	99.94999
9500	97.46794	10000	100.00000

# APPENDIX 15: Sums and Sums of Squares of Natural Numbers

FIRST 50 NATURAL NUMBERS

Natural number	Sum	Sum of squares
1	1	1
2	3	5
3	6	14
4	10	30
5	15	55
6	21	91
7	28	140
8	36	204
9	45	285
10	55	385
11	66	506
12	78	650
13	91	819
14	105	1 015
15	120	1 240
16	136	1 496
17	153	1 785
18	171	2 109
19	190	2 470
20	210	2 870
21	231	3 311
22	253	3 795
23	276	4 324
24	300	4 900
25	325	5 525
26	351	6 201
27	378	6 930
28	406	7 714
29	435	8 555
30	465	9 455
31	496	10 416
32	528	11 440
33	561	12 529
34	595	13 685
35	630	14 910
36	666	16 206
37	703	17 575
38	741	19 019
39	780	20 640
40	820	22 140
41	861	23 821
42	903	25 585
43	946	27 434
44	990	29 370
45	1 035	31 395
46	1 081	33 511
47	1 128	35 720
48	1 176	38 024
49	1 225	40 425
50	1 275	42 925

FIRST 50 ODD NATURAL NUMBERS

Odd natural number	Sum	Sum of squares
1	1	1
3	4	10
5	9	35
7	16	84
9	25	165
11	36	286
13	49	455
15	64	680
17	81	969
19	100	1 330
21	121	1 771
23	144	2 300
25	169	2 925
27	196	3 654
29	225	4 495
31	256	5 456
33	289	6 545
35	324	7 770
37	361	9 139
39	400	10 660
41	441	12 341
43	484	14 190
45	529	16 215
47	576	18 424
49	625	20 825
51	676	23 426
53	729	26 235
55	784	29 260
57	841	32 509
59	900	35 990
61	961	39 711
63	1 024	43 680
65	1 089	47 905
67	1 156	52 394
69	1 225	57 155
71	1 296	62 196
73	1 369	67 525
75	1 444	73 150
77	1 521	79 079
79	1 600	85 320
81	1 681	91 881
83	1 764	98 770
85	1 849	105 995
87	1 936	113 564
89	2 025	121 485
91	2 116	129 766
93	2 209	138 415
95	2 304	147 440
97	2 401	156 849
99	2 500	166 650

## APPENDIX 16: Flexible Calendar of Working Days

### CALENDAR DAYS, SUNDAYS, SATURDAYS, AND HOLIDAYS, BY MONTHS, 1898-1976

There are 14 distinct calendar patterns, referred to in the calendar by code number. In the code table below the years are arranged consecutively within columns. Any year can be located by reading down the proper column. Then read across to ascertain the code number. For instance, 1945 is located by reading down in the fifth column, and the code number is seen to be IV. Row IV of the calendar gives information concerning 1945, as well as concerning 1900, 1906, 1917, 1923, 1934, 1951, 1962, and 1973.

CODE TABLE

Year								Code number
1898	1910 <sup>fe</sup>	1921 <sup>fe</sup>	1927	1938	1949	1955	1966	I
...	...	...	1928*	...	...	1956* <sup>f</sup>	...	II
1899 <sup>f</sup>	1911	1922	...	1939	1950	...	1967 <sup>fe</sup>	III
1900†	...	1923 <sup>f</sup>	...	...	1951 <sup>fe</sup>	...	...	IV
...	1912*	...	...	1940* <sup>fe</sup>	...	...	1968*	V
1901	...	...	1929 <sup>fe</sup>	...	...	1957	...	VI
1902 <sup>fe</sup>	1913 <sup>fe</sup>	...	1930	1941	...	1958	1969	VII
...	...	1924*	...	...	1952*	...	...	VIII
1903	1914	1925	1931	1942	1953	1959 <sup>fe</sup>	1970 <sup>fe</sup>	IX
...	1915	1926	...	1943	1954	...	1971	X
1904*	...	...	1932* <sup>fe</sup>	...	...	1960*	...	XI
...	1916*	...	...	1944*	...	...	1972* <sup>f</sup>	XII
1905	...	...	1933	...	...	1961 <sup>f</sup>	...	III
1906	1917	...	1934 <sup>f</sup>	1945 <sup>f</sup>	...	1962	1973	IV
1907 <sup>fe</sup>	1918 <sup>fe</sup>	...	1935	1946	...	1963	1974	VI
1908*	...	...	1936*	...	...	1964* <sup>fe</sup>	...	XIII
...	1919	...	...	1947	...	...	1975 <sup>fe</sup>	VII
1909	...	...	1937 <sup>fe</sup>	...	...	1965	...	X
...	1920*	...	...	1948* <sup>fe</sup>	...	...	1976*	XIV

\* Leap Year; February has 29 days.

† 1900 was not a Leap Year.

<sup>f</sup> Good Friday occurred in March.

<sup>e</sup> Easter occurred in March.



## CALENDAR

The first row for each year gives the number of Sundays in parentheses ( ) and Saturdays in brackets [ ] in each month. The second row shows the occurrence of holidays. Holidays occurring on Sundays are enclosed in parentheses; those on Saturdays are enclosed in brackets. For information concerning the states in which specific holidays are observed, see *The World Almanac and Book of Facts* (published annually by the New York World-Telegram and Sun, New York City).

Following is a key to the symbols used on the calendar:

N	New Year's Day—January 1.	D	Labor Day—First Monday in September.
L	Lincoln's Birthday—February 12.	C	Columbus Day—October 12.
W	Washington's Birthday—February 22.	V	Election Day—First Tuesday after First Monday in November.
F	Good Friday.	A	Veteran's Day—November 11 (beginning 1918).
E	Easter.	T	Thanksgiving Day.
M	Memorial Day—May 30.	X	Christmas Day—December 25.
J	Independence Day—July 4.		

Code number	Jan 31	Feb 28	Mar 31	Apr 30	May 31	Jun 30	Jul 31	Aug 31	Sep 30	Oct 31	Nov 30	Dec 31
I	(5) [5] [N]	(4) [4] [L] W	(4) [4]	(4) [5] F (E)	(5) [4] M	(4) [4]	(5) [5] J	(4) [4]	(4) [4] D	(5) [5] (C)	(4) [4] V A T	(4) [5] (X)
II	(5) [4] (N)	(4) [4] L W	(4) [5]	(5) [4] F (E)	(4) [4] M	(4) [5]	(5) [4] J	(4) [4]	(5) [5] D	(4) [4] C	(4) [4] V (A) T	(5) [5] X
III	(5) [4] (N)	(4) [4] (L) W	(4) [4]	(5) [5] F (E)	(4) [4] M	(4) [4]	(5) [5]	(4) [4]	(4) [5]	(5) [4] C	(4) [4] V (A) T	(5) [5] X
IV	(4) [4] N	(4) [4] L W	(4) [5]	(5) [4] F (E)	(4) [4] M	(4) [5]	(5) [4] J	(4) [4]	(5) [5] D	(4) [4] C	(4) [4] V (A) T	(5) [5] X
V	(4) [4] N	(4) [4] L W	(5) [5]	(4) [4] F (E)	(4) [4] M	(5) [5]	(4) [5] J	(4) [5]	(5) [4] D	(4) [4] [C]	(4) [5] V A T	(5) [4] X
VI	(4) [4] N	(4) [4] L W	(5) [5]	(4) [4] F (E)	(4) [4] M	(5) [5]	(4) [4] J	(4) [5]	(5) [4] D	(4) [4] [C]	(4) [5] V A T	(5) [4] X
VII	(4) [4] N	(4) [4] L [W]	(5) [5]	(4) [4] F (E)	(4) [5] M	(5) [4]	(4) [4] J	(5) [5]	(4) [4] D	(4) [4] (C)	(4) [5] V A T	(4) [4] X
VIII	(4) [4] N	(4) [4] L W	(5) [5]	(4) [4] F (E)	(4) [5] M	(5) [4]	(4) [4] J	(5) [5]	(4) [4] D	(4) [4] (C)	(5) [5] V A T	(4) [4] X
IX	(4) [5] N	(4) [4] L (W)	(5) [4]	(4) [4] F (E)	(5) [5] [M]	(4) [4]	(4) [4] [J]	(5) [5]	(4) [4] D	(4) [5] G	(5) [4] V A T	(4) [4] X
X	(5) [5] N	(4) [4] L W	(4) [4]	(4) [4] F (E)	(5) [5] (M)	(4) [4]	(4) [5] (J)	(5) [4]	(4) [4] D	(5) [5] C	(4) [4] V A T	(4) [4] [X]
XI	(5) [5] N	(4) [4] L W	(4) [4]	(4) [5] F (E)	(5) [4] M	(4) [4]	(5) [5] J	(4) [4]	(4) [4] D	(5) [5] C	(4) [4] V A T	(4) [5] (X)
XII	(5) [5] N	(4) [4] [L] W	(4) [4]	(5) [5] F (E)	(4) [4] M	(4) [4]	(5) [5] J	(4) [4]	(4) [5] D	(5) [4] C	(4) [4] V A T	(5) [5] X
III	(5) [4] (N)	(4) [4] (L) W	(4) [4]	(5) [5] F (E)	(4) [4] M	(4) [4]	(5) [5] J	(4) [4]	(4) [5] D	(5) [4] C	(4) [4] V (A) T	(5) [5] X
IV	(4) [4] N	(4) [4] L W	(4) [5]	(5) [4] F (E)	(4) [4] M	(4) [5]	(5) [4] J	(4) [4]	(5) [5] D	(4) [4] C	(4) [4] V (A) T	(5) [5] X
VI	(4) [4] N	(4) [4] L W	(5) [5]	(4) [4] F (E)	(4) [4] M	(5) [5]	(4) [4] J	(4) [5]	(5) [4] D	(4) [4] [C]	(4) [5] V A T	(5) [4] X
XIII	(4) [4] N	(4) [5] L [W]	(5) [4]	(4) [4] F (E)	(5) [5] [M]	(4) [4]	(4) [4] [J]	(5) [5]	(4) [4] D	(4) [5] C	(5) [4] V A T	(4) [4] X
VII	(4) [4] N	(4) [4] L [W]	(5) [5]	(4) [4] F (E)	(4) [5] M	(5) [4]	(4) [4] J	(5) [5]	(4) [4] D	(4) [4] (C)	(5) [5] V A T	(4) [4] X
X	(5) [5] N	(4) [4] L W	(4) [4]	(4) [4] F (E)	(5) [5] (M)	(4) [4]	(4) [5] (J)	(5) [4]	(4) [4] D	(5) [5] C	(4) [4] V A T	(4) [4] [X]
XIV	(4) [5] N	(5) [4] L (W)	(4) [4]	(4) [4] F (E)	(5) [5] (M)	(4) [4]	(4) [5] J	(5) [4]	(4) [4] D	(5) [5] C	(4) [4] V A T	(4) [4] [X]

## APPENDIX 17: A Review of Some Topics in Elementary Mathematics

### A.1 THE SUMMATION OPERATOR $\Sigma$

The following rules and definitions apply to the summation operator  $\Sigma$ .

$$(1) \quad \sum_{i=1}^n X_i = X_1 + X_2 + \cdots + X_n$$

and is often written  $\Sigma X$  when the upper and lower limits of summation are understood.

$$(2) \quad \Sigma X^2 = X_1^2 + X_2^2 + \cdots + X_n^2$$

which is *not* the same as

$$(3) \quad (\Sigma X)^2 = (X_1 + X_2 + \cdots + X_n)^2.$$

$$(4) \quad \Sigma (X + Y) = \Sigma X + \Sigma Y$$

$$(5) \quad \Sigma X \Sigma Y = (\Sigma X)(\Sigma Y)$$

$$(6) \quad \sum_{i=1}^n k = nk$$

where  $k$  is any constant.

$$(7) \quad \sum_{i=1}^n kX_i = k \sum_{i=1}^n X_i$$

where  $k$  is any constant and  $X_i$  is a variable.

$$(8) \quad \sum_{i=1}^m \sum_{j=1}^n X_{ij} = X_{11} + X_{12} + \cdots + X_{1n} + X_{21} + X_{22} \\ + \cdots + X_{2n} + \cdots + X_{m1} + X_{m2} + \cdots + X_{mn}$$

$$(9) \quad \Sigma X^a = \Sigma (X^b)^c = \Sigma (X^d)(X^e)$$

where  $a = bc = d + e$ .

### A.2 FACTORIAL AND COMBINATORIAL NOTATION

The exclamation point after a symbol or number is read "factorial."

$$(1) \quad X! = X(X-1)(X-2)\cdots(1)$$

$$(2) \quad 1! = 1$$

$$(3) \quad 0! = 1$$

The symbol  $\binom{n}{d}$  means the number of combinations of  $n$  things taken  $d$  at a time.

$$(4) \quad \binom{n}{d} = \frac{n!}{d! (n-d)!}$$

$$(5) \quad \binom{n}{d} = \binom{n}{n-d}$$

$$(6) \quad \binom{n}{n} = \binom{n}{0} = 1$$

$$(7) \quad \binom{n}{n-1} = \binom{n}{1} = n$$

### A.3 RULES OF ROUNDING

(1) If the leftmost of the digits to be discarded is less than 5, the preceding digit is not affected. Thus:

117.746 becomes 117.7

when rounded to four digits.

(2) If the leftmost of the digits to be discarded is greater than 5, or is 5 followed by digits not all of which are zero if carried to a sufficient number of digits, the preceding digit is increased by one. Thus:

129.673 becomes 129.7

when rounded to four digits and

87.22500001 becomes 87.23

(3) If the leftmost of the digits to be discarded is exactly 5, followed by zeros only, the preceding digit is increased by one if it is odd, and left unchanged if it is even. Thus:

103.55 becomes 103.6 and 103.45 becomes 103.4

when rounded to four digits.

Generally speaking, a *final* answer that is the result of the multiplication or division of numbers that have been rounded cannot have more significant digits than the original number with the fewest significant digits. For example,

$$2.16/2.5 = 0.864$$

but the final answer should be rounded to 0.86 or perhaps 0.9 since there are only two significant digits in the denominator of the fraction. In a similar manner,

$$2.3(2.16) = 4.968$$

should be rounded to 5.0.

A conservative rule is to record one less digit in a final answer than in the original number with the fewest significant digits.

#### A.4 OTHER SYMBOLS

The following symbols are used in this book and are not explained elsewhere.

<i>Symbol</i>	<i>Meaning</i>
$a \neq b$	$a$ is not equal to $b$
$a < b$	$a$ is less than $b$
$a \leq b$	$a$ is less than or equal to $b$
$a > b$	$a$ is greater than $b$
$a \geq b$	$a$ is greater than or equal to $b$
$a \doteq b$	$a$ is approximately equal to $b$



# Index

## A

Acceptance number, 164, 190  
Acceptance region, 130  
Accounting, use of statistics in, 4  
Accuracy, 107  
Adjusted frequencies, 15  
Alienation, coefficient of, 209n  
Alpha:  
    Bayesian viewpoint, 190–191  
    classical advice on setting, 141  
Amplitude adjusted index, 378  
Analysis of variance:  
    one basis of classification:  
        case of two columns, 271  
        computation, 269–271  
        model, 267  
        symbolic statement of test, 267–268  
        unequal sample sizes, 271  
    two bases of classification:  
        case of two columns, 276  
        computation, 275–276  
        model, 272–274  
        symbolic statement of test, 274  
Arithmetic mean:  
    algebraic manipulation of, 31  
    computation of:  
        grouped data, 32–33

        weighted, 23  
confidence limits (*see* Confidence limits)  
effect of kurtosis, 31  
effect of open-end classes, 32  
effect of skewness and extreme values, 29–31  
hypotheses concerning (*see* Hypothesis tests)  
mechanical meaning, 29  
modified, 27n, 355  
standard error of (*see* Standard error of various statistics)  
sum of deviations from, 24  
sum of squares of deviations from, 24–25  
Array, 8–9  
Autocorrelation, 326, 369, 375–376  
Autoregressive model, 375  
Average deviation (*see* Mean deviation)

## B

Barton, H.C., Jr., 359n  
Bayes, T. (*see* Probability)  
Bernoulli trials, 72  
Beta coefficients, 240–241

Bias, 40, 44, 106  
 Binomial coefficients, table of, 401  
 Binomial distribution, 72-76, 82-87,  
     162-164, 168-169  
 Bivariate distribution, 206, 210  
 BLUE estimator, 264n  
 BMD, Biomedical Computer Programs,  
     247n  
 Bolch, Ben W., 405  
 Bradley, Ralph Allen, 157

## C

Calendar variation:  
     length of month variation, 349  
     trading day variation, 349-350  
 Causal models, 382  
 Central limit theorem (*see* Normal  
     distribution)  
 Chebyshev's inequality, 93  
 Chi square:  
     contingency, 279-280  
     continuity correction, 280-281  
     defined, 171  
     goodness of fit, 99, 276-277, 368  
 Class boundary (*see* Class limits)  
 Class interval, 10  
 Class limits:  
     actual, 10  
     stated, 10  
 Clopper, J.C., 168  
 Cochran, William G., 120n  
 Column diagram, 11  
 Combinatorial notation, 433-434  
 Confidence limits:  
     arithmetic mean with specified  
         population variance, 139-141  
     arithmetic mean with unspecified  
         population variance, 153-154  
     correlation coefficients:  
         multiple, 242n  
         partial, 245  
         simple, 219-220  
     difference between means, 156  
     intercept of simple regression line, 204  
     mean difference, 158  
     proportions, 167-169  
         charts for confidence limits for  
             proportions, 403-404  
     regression:  
         coefficients, 204, 261-262  
         line, 205  
         prediction, 205  
     sample size determination, 145-146  
     slope of simple regression line, 204  
     standard deviations, 172

Conner, William J., 266  
 Consumer's risk, 164n  
 Continuity correction, 165n, 280-281  
 Continuous variables, 7  
 Control charts:  
     defects, 179  
     means, 178, 179-184  
     proportion defective, 179  
     ranges, 178  
         table for setting control limits for  
             ranges, 405  
 Control limits, 178  
 Correlation (multiple):  
     adjusted for degrees of freedom, 247  
     computation, 235, 239n  
     confidence limits (*see* Confidence  
         limits)  
     hypotheses concerning (*see* Hypothesis  
         tests)  
     interpretation, 235-239  
 Correlation (multiple-partial), 246  
 Correlation (partial):  
     computation, 236, 240  
     confidence limits (*see* Confidence  
         limits)  
     hypotheses concerning (*see* Hypothesis  
         tests)  
     interpretation, 235-239  
 Correlation (simple):  
     adjusted for degrees of freedom, 227  
     alternative concepts, 209-213, 226-227  
     causation, 215-216  
     computation (*see* alternative concepts)  
     confidence limits (*see* Confidence  
         limits)  
     hypotheses concerning (*see* Hypothesis  
         tests)  
     interpretation, 213-216  
     ranked data, 222-223  
     time series, 312, 370-380  
 Correlogram, 376  
 Cowden, Dudley J., 33n, 82n, 172n,  
     240n, 246n, 263n, 266, 326n,  
     365n, 401n, 408n  
 Cowden, Mercedes S., 408n  
 Critical Region (*see* Rejection region)  
 Croxton, Frederick E., 33n, 326n  
 Cyclical movements, 363-367

## D

David, F.N., 219n  
 Defectives, 74, 77  
 Defects, 76, 77  
 Degrees of freedom, 105, 148-149

Dependent variable, 206  
 Determination, coefficient of, 209  
 Difference operator, 318n  
 Diffusion index, 377-378  
 Discrete variables, 7  
 Distributed lag model, 379n  
 Distribution-free statistics, 223  
 Dixon, W.J., 247n  
 Doolittle Method, abbreviated, 236n,  
 262-263  
 Dummy variables, 351n  
 Duncan, Acheson J., 183n

## E

e:  
   defined, 77, 85n  
   table of negative exponents, 402  
 Error, types of, 111  
 Errors of estimate, 195  
 Explained variable (*see* Dependent  
   variable)  
 Exponential (*see* Growth curves)

## F

F distribution:  
   defined, 173  
   probability table, 395-397  
   ratios following, 173, 217, 241-242,  
   266, 269, 270, 274  
 Factorial notation, 433  
 Faddeeva, V.N., 253n  
 Fieller, E.C., 211n  
 Finance, use of statistics in, 4  
 Fisher, Irving, 305  
 Fisher, R.A., 58, 125n, 219n, 394n, 395,  
   398  
 Fisher's k-statistics, 58  
 Forecasting methods, 370  
 Frazier, Dale, 272  
 Freeman, William W.K., 4  
 Frequency distribution:  
   cumulative, 18-20  
   graphic presentation, 11-13  
   interpretation of, 16-18  
   percentage, 18  
 Frequency polygon, 11  
 Fry, Thornton C., 401n

## G

Gauss-Markov Theorem, 263-264  
 Geometric Mean, 34-36

Gompertz (*see* Growth curves)  
 Goodness of fit, 276-279  
 Gosset, W.S., 148  
 Gram-Charlier System of Frequency  
   Curves, 82  
 Graybill, Franklin A., 264n  
 Greenhouse, Samuel W., 280n  
 Grizzle, James E., 280  
 Growth curves:  
   exponential, 327-329  
   Gompertz, 336-337  
   logistic, 337-338  
   modified exponential, 332-336  
   second degree exponential, 329-332  
 Guilford, J.P., 213n

## H

Harmonic mean, 36  
 Hartley, H.O., 99n, 147n, 281n, 326n,  
   395, 398, 404n, 405n  
 Heterogeneous data, 28  
 Histogram (*see* Column diagram)  
 Hotelling, Harold, 219n, 221  
 Hypergeometric distribution, 78, 166,  
   186-187  
 Hypothesis testing, defined, 110  
 Hypothesis tests:  
   arithmetic mean:  
     difference between, 154-156  
     equality of several (*see* Analysis of  
       variance)  
     mean difference, 156-158  
     sample size determination, 143-145  
     single mean (specified population  
       variance), 127-132  
     single mean (unspecified population  
       variance), 147-153  
   correlation coefficients:  
     difference between, different  
       populations, 220-221  
     difference between, same population,  
       221-222  
     significance:  
       multiple, 241  
       partial, 242-244  
       simple, 216-218  
       table for significance test, 406  
   moments, 99  
   proportions:  
     difference between, 170-171  
     single proportion, finite population,  
       166-167  
     single proportion, infinite population,  
       162-166



- regression coefficients:
  - partial, 261-262
  - simple, 202-204
- standard deviation:
  - difference between, 172-174
  - single standard deviation, 171-172

## I

- IBM scientific subroutine package, 326n
- Identification, 384
- Independent variable, 206
- Index numbers:
  - base selection, 286
  - chain, 303
  - consumer price index, 285, 295
  - data selection, 287
  - Fisher's Ideal Index, 296, 298, 303, 305
  - Laspeyres' Formula, 293-296
  - Marshall-Edgeworth Formula, 296
  - Paasche's Formula, 293-296
  - price relatives, 291, 298-301
  - quantity relatives, 291, 298-301
  - shifting base, 301-302
  - splicing, 302-303
  - Stuvel's Formula, 306
- tests:
  - circular, 304n
  - factor reversal, 305
  - proportionality, 305
  - time reversal, 304
- weight selection, 287
- wholesale price index, 295n
- Indirect least squares, 384
- Interdependent model, 383
- Interval estimation, defined, 109-110
- Irregular movements, 367-369

## J

- Johnston, J., 198n, 375n
- Jointly determined variables, 383

## K

- Kelley, Truman L., 13n
- Kendall's tau, 223n
- Keynes, J.M., 63
- Kinsey, Alfred C., 5
- Klein, Sidney, 33
- Koyck, L.M., 379n
- Kramer, K.H., 242n

## Kurtosis:

- graphic interpretation, 17
- measure of:
  - Geary's, 59
  - moments, 55
- standard error of measure of relative kurtosis (*see* Standard error of various statistics)

## L

- Latané, Henry A., 359n
- Latin square, 124
- Least squares criterion, 195
- Leptokurtic curve, 17
- Lewis, T., 211n
- Logarithms, rules for elementary operations, 409
- Logarithms, table of common, 409-413
- Logistic (*see* Growth curves)
- Loss:
  - average, 181
  - conditional, 180
  - expected average, 181
  - opportunity, 184
  - total expected, 181
- Lovell, Michael C., 351n, 359n

## M

- Mantel, Nathan, 280n
- Marketing, use of statistics in, 3
- Matrix algebra:
  - addition and subtraction, 250
  - adjoint, 255
  - cofactors, 255
  - determinant, 253-254
  - identity (*see* unit)
  - inversion, 253-256
  - matrix multiplication, 250-252
  - minor, 254
  - nonsingular, 254
  - scalar multiplication, 250
  - singular, 254
  - symmetric, 252
  - transposition, 252
  - unit, 252-253
- vectors:
  - column, 256
  - row, 256
- Maximum likelihood, 44n, 313n
- Mean deviation, 42
- Median:
  - characteristics, 20, 28-33

computation from grouped data, 33  
 effect of kurtosis, 31  
 effect of open-end classes, 32  
 effect of skewness, 29  
 Mesokurtic curve, 18  
 Mid-range, 26, 42  
 Mid-value, 10  
 Minimum variance unbiased estimators, 107  
 Mode:  
   characteristics, 27  
   computation from grouped data, 28  
   effect of kurtosis, 31  
   effect of open-end classes, 32  
   effect of skewness, 30  
 Modified exponential (*see* Growth curves)  
 Molina, E.C., 78n  
 Moments:  
   computation, 56–58  
   definition, 54–55  
   hypotheses concerning (*see* Hypothesis tests)  
   use of in measuring skewness and kurtosis, 55–56  
 Moore, Geoffrey H., 378n  
 Moving arcs (*see* Moving average, polynomially weighted)  
 Moving average:  
   as trend, 338–342, 363–364  
   weighted:  
     binomial, 365  
     Henderson, 365n  
     polynomial, 365–367  
  
**N**  
 Nerlove, Marc, 359n  
 Nondetermination, coefficient of, 209n  
 Nonparametric statistics (*see* Distribution free statistics)  
 Normal distribution:  
   areas under, 89–92  
   central limit theorem, 88  
   defined, 85  
   fitting to a frequency distribution, 93–99  
   as limit of binomial distribution, 82–87  
   probability tables, 391–393  
 Normal equations, 196, 231, 257, 313, 319, 323  
 Number defective:  
   confidence limits (*see* Confidence limits)  
   expected value, 106, 160

hypotheses concerning (*see* Hypothesis tests)  
 variance:  
   finite population, 161  
   infinite population, 109, 161

## O

Ogive, 18–20  
 Open-end classes, 15  
 Operating characteristic function, 133n, 181  
 Orthogonal polynomials, 326–327

## P

Partition values, 20  
 Peach, Paul, 62n  
 Pearl-Read curve (*see* Logistic growth curve)  
 Pearson, E.S., 99n, 147n, 168, 211n, 281n, 326n, 395, 398, 404n, 405n  
 Pearson, K., 59  
 Pearsonian system of frequency curves, 56, 57, 82  
 Percentile, 20  
 Personnel administration:  
   use of statistics in, 3  
 Pilot study, 121  
 Platykurtic curve, 17  
 Point estimation:  
   mean, 102  
   number defective, 106  
   proportion defective, 106  
   variance, 102  
 Poisson distribution, 76–78, 164  
 Polynomials (*see* Trend and moving average, weighted)  
 Population:  
   defined, 1  
   finite, 102  
   infinite, 102  
 Power of a test:  
   arithmetic mean, 132–139, 144, 147n, 181  
   F test, 174n  
   number defective, 190  
   proportions, 163, 164n  
 Precision, 107  
 Predetermined variables, 383  
 Predictand (*see* Dependent variable)  
 Predictor (*see* Independent variable)  
 Probability:  
   addition rule, 67–68

- axioms, 64–65
- Bayes' theorem, 78–79, 188–189
- conditional, 66–67
- definitions, 61–63
- density, 85
- dependent events, 66
- independent events, 65
- joint, 67
- marginal, 67
- multiplication rule, 68–69
- mutually exclusive events, 65
- paper, 97–98
- prior, 179
- sample space, 64
- tree, 69
- Venn diagram, 65
- Producer's risk, 162, 164n
- Production, use of statistics in, 3
- Proportions:
  - confidence limits (*see* Confidence limits)
  - expected value, 106, 160
  - hypotheses concerning (*see* Hypothesis tests)
  - variance:
    - finite population, 161
    - infinite population, 109, 161

## Q

- Quadratic mean, 36–37
- Qualitative variables, 7
- Quality control, 176–179
- Quantitative variables, 7
- Quartile, 20
- Quartile deviation, 42n

## R

- Random numbers:
  - table, 408
  - use, 117–118
- Random variable, 70
- Range:
  - correction for bias, 40
  - table for use in correcting for bias, 405
  - definition, 40
- Raw data, 7–8
- Reciprocals, table of, 414–429
- Reduced form equations, 383–384
- Reflection, 239

- Regressand (*see* Dependent variable)
- Regression (multiple):
  - components of variation, 234
  - computation, 228–234
  - confidence limits (*see* Confidence limits)
  - Doolittle method, 236n, 262–263
  - hypotheses concerning (*see* Hypothesis tests)
  - matrix algebra approach, 257–264
  - transformed data, 245–246
- Regression (simple):
  - components of variation, 200–202
  - computation, 192–197
  - confidence limits (*see* Confidence limits)
  - hypotheses concerning (*see* Hypothesis tests)
- Regressor (*see* Independent variable)
- Regret, 184
- Rejection number, 164, 190
- Rejection region, 130
- Relative dispersion, 48–49
- Reliability, 107
- Robust test, 267
- Romig, Harry G., 76n, 168
- Root mean square deviation, 44
- Rounding, 434–435

## S

- St. Petersburg Paradox, 191
- Salzman, Lawrence, 350n
- Sampling:
  - absolute vs. relative sample size, 115
  - acceptance, 184
  - designs:
    - area, 123
    - cluster, 121
    - multistage, 122
    - random point, 124
    - sequential, 124
    - simple, 119–120
    - stratified, 120–121
  - exercise of judgment, 115
  - random, 116
  - systematic, 118
- Savage, L.J., 63
- Scatter diagram, 194
- Seasonal adjustment:
  - ratio to moving average:
    - moving, 359–363
    - stable, 350–359

- regression, 351n
  - seasonal adjustment factor, 355
  - Second degree exponential (*see* Growth curves)
  - Serial correlation, 375n
  - Shartel, C.L., 230n
  - Shiskin, Julius, 378n
  - Siegel, Sidney, 223n
  - Skewness:
    - graphic interpretation, 17, 52, 53
    - measure of:
      - Fisher's *k* statistics, 58
      - moments, 55
      - K. Pearson's measure, 59
    - standard error of measure of relative skewness (*see* Standard error of various statistics)
  - Snedecor, George W., 13n
  - Spearman's coefficient, 223
  - Spectral analysis, 359
  - Squares and square roots, table of, 414-429
  - Standard deviation, *s*:
    - computation:
      - grouped data, 47-48
      - ungrouped data, 46-47
    - confidence limits (*see* Confidence limits)
    - correction for bias, 45
    - table for use in correcting for bias, 405
    - definition, 45
    - hypotheses concerning (*see* Hypothesis tests)
    - mathematical properties, 45-46
  - Standard deviation, SD:
    - computation:
      - grouped data, 47-48
      - ungrouped data, 43, 46-47
    - correction for bias, 44
    - table for use in correcting for bias, 405
    - definition, 43
    - mathematical properties, 45-46
  - Standard error of estimate, 208, 234, 261
  - Standard error of regression (*see* Standard error of estimate)
  - Standard error of various statistics:
    - arithmetic mean, 108-109
    - correlation coefficients
      - (*z* transformation):
        - difference between, 220
        - simple, 219
        - partial, 245
    - difference between two means, 155
    - difference between two proportions, 170
    - estimate (regression), 208, 234, 261
    - Fisher's *k*-statistic measure of relative kurtosis, 58n
    - Fisher's *k*-statistic measure of relative skewness, 58n
    - intercept of simple regression line, 202
    - mean difference, 157
    - moment measure of relative kurtosis, 58n, 99
    - moment measure of relative skewness, 58n, 99
    - multiple regression coefficients, 261
    - slope of simple regression line, 202
  - Standardized death rates, 54n
  - Standardized distributions, 51-54
  - Standard partial regression coefficient (*see* Beta coefficients)
  - Statistic, distinguished from parameter, 55, 101
  - Statistics:
    - defined, 1
    - descriptive, 2
    - inferential, 2
  - Stepwise regression, 247
  - Structural equations, 383
  - Student's distribution (*see* *t* distribution)
  - Sturges' rule, 13n
  - Summation operator, 433
  - Sums and sums of squares, table of, 430
- T**
- t* distribution:
    - compared to normal distribution, 149
    - measure of kurtosis, 149n
    - probability table, 394
    - variance, 149n
  - Theil, H., 384n
  - Theil's coefficient, 384-387
  - Tiku, Moti Lal, 174n
  - Time series model, 308
  - Tinbergen arrow diagram, 382
  - Trend (*see also* Growth curves and moving averages):
    - changing units and shifting origin, 317-318
    - methods of fitting, 313n
    - polynomial:
      - first degree, 313-317
      - second degree, 318-322
      - third degree, 322-325
    - selection of type and period for fit, 342-345

test for significance, 312  
 Tschuprow's coefficient, 283  
 Twain, Mark, 6

## U

Unbiased estimators:  
   defined, 104  
   list of, 106  
 Unit distributions, 52

## V

Variance, 43  
 Variance ratio (*see* F distribution)  
 Variate (*see* Random variable)  
 Variation:  
   additive property, 200, 234, 242, 268,  
     274  
   computation, 43  
   diagrammatic representation in  
     multiple regression, 237

## W

Width of class interval (*see* Class  
   interval)  
 Wilks, S.S., 62n  
 Wold, Herman O.A., 384n  
 Working days, table of, 431–432  
 Wyer, Rolf, 225

## Y

Yates, F., 125n, 278, 394n, 395, 398

## Z

z score, 54  
 z transformation (correlation  
   coefficients), 219–222, 245  
   table for calculation of  
     z transformation, 407

